DE GRUYTER

Viola Wiegand,
Michaela Mahlberg (Eds.)

CORPUS LINGUISTICS, CONTEXT AND CULTURE

DISKURSMUSTER DISCOURSE PATTERNS

Corpus Linguistics, Context and Culture

Diskursmuster
Discourse Patterns

Edited by
Beatrix Busse and Ingo H. Warnke

Volume 15

Corpus Linguistics, Context and Culture

Edited by
Viola Wiegand and Michaela Mahlberg

DE GRUYTER

ISBN 978-3-11-048672-8
e-ISBN (PDF) 978-3-11-048907-1
e-ISBN (EPUB) 978-3-11-048711-4

Library of Congress Control Number: 2018958738

Bibliografische Information der Deutschen Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Part I: **Discourse contexts and cultures**

Part II: **Contexts of lexis and grammar**

Part III: **Learner contexts**

Viola Wiegand and Michaela Mahlberg

Introduction

On context and culture in corpus linguistics

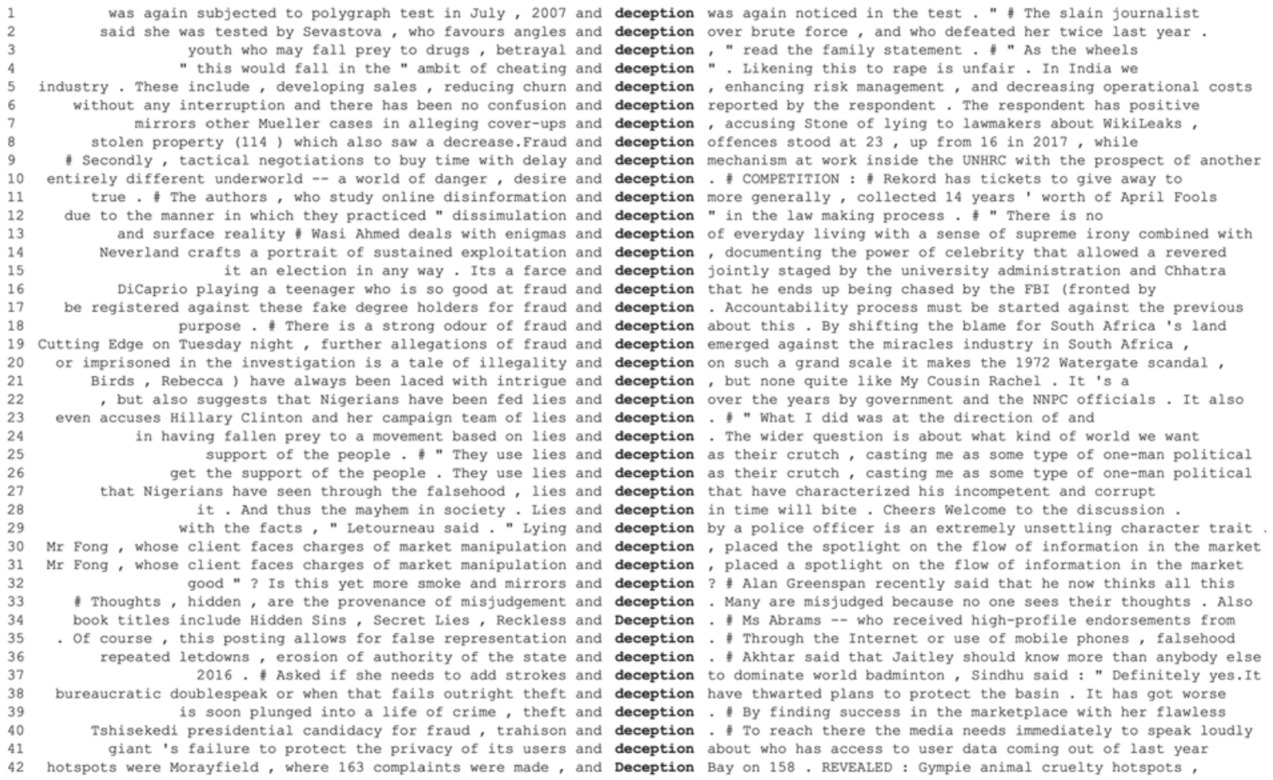
Corpora provide invaluable insights into language in use. The texts in corpora capture evidence of social interactions and behaviours. Hence, in corpus linguistics, ‘context’ is a key concept. In very practical terms, most general corpus software packages will have KWIC (“key word in context”) display options. The KWIC display played a crucial part in the “corpus revolution” in the 1980s when the systematic observation of lexico-grammatical patterns fundamentally changed approaches in lexicography (Sinclair 1987). As corpus linguistics developed, the concordance was the central tool that enabled a fresh view on words and their meanings, recognising the crucial role of collocation and phraseology.

In recent work in corpus linguistics, it seems the popularity of the KWIC display format has been declining, while large-scale statistical analysis methods have become much more common. In practice, concordance analysis is still part of the methodology of many studies, but by now, its role appears almost so fundamental that it is implied and not necessarily made explicit anymore. While several contributions to this volume feature findings generated with the help of concordance analyses, the chapters quote examples from concordance lines rather than display actual concordances. It is worth taking a look at a concordance sample to highlight how the connection between context and culture becomes visible in this display format.

The word *deception* characterises a crucial phenomenon of twenty-first century society. Figure 1 shows a sample of concordance lines for *deception* in the NOW corpus. The NOW corpus (News on the Web; Davies 2013) contains several billion words of data from web-based newspapers and magazines from 2010 to the present time. It is a dynamic corpus that grows steadily about 140-160 million words per month. We have selected lines in which *deception* is preceded by *and* to form part of a list. In this sample, *deception* shows a tendency to occur with words that have meanings relating to dishonesty, including *betrayal* (line 3), *cheating* (line 4), *cover-ups* (line 7), *disinformation* (line 11), *farce* (line 15), *fraud* (lines 16—19), *illegality* (line 20), *lies* (lines 22—28), *lying* (line 29), *manipulation*

Viola Wiegand, University of Birmingham, v.wiegand@bham.ac.uk
Michaela Mahlberg, University of Birmingham, m.a.mahlberg@bham.ac.uk

https://doi.org/10.1515/9783110489071-001



```
1          was again subjected to polygraph test in July , 2007 and deception was again noticed in the test . " # The slain journalist
2         said she was tested by Sevastova , who favours angles and deception over brute force , and who defeated her twice last year .
3                   youth who may fall prey to drugs , betrayal and deception , " read the family statement . # " As the wheels
4                  " this would fall in the " ambit of cheating and deception " . Likening this to rape is unfair . In India we
5  industry . These include , developing sales , reducing churn and deception , enhancing risk management , and decreasing operational costs
6      without any interruption and there has been no confusion and deception reported by the respondent . The respondent has positive
7              mirrors other Mueller cases in alleging cover-ups and deception , accusing Stone of lying to lawmakers about WikiLeaks ,
8        stolen property (114 ) which also saw a decrease.Fraud and deception offences stood at 23 , up from 16 in 2017 , while
9     # Secondly , tactical negotiations to buy time with delay and deception mechanism at work inside the UNHRC with the prospect of another
10  entirely different underworld -- a world of danger , desire and deception . # COMPETITION : # Rekord has tickets to give away to
11        true . # The authors , who study online disinformation and deception more generally , collected 14 years ' worth of April Fools
12     due to the manner in which they practiced " dissimulation and deception " in the law making process . # " There is no
13          and surface reality # Wasi Ahmed deals with enigmas and deception of everyday living with a sense of supreme irony combined with
14        Neverland crafts a portrait of sustained exploitation and deception , documenting the power of celebrity that allowed a revered
15                    it an election in any way . Its a farce and deception jointly staged by the university administration and Chhatra
16         DiCaprio playing a teenager who is so good at fraud and deception that he ends up being chased by the FBI (fronted by
17     be registered against these fake degree holders for fraud and deception . Accountability process must be started against the previous
18                purpose . # There is a strong odour of fraud and deception about this . By shifting the blame for South Africa 's land
19 Cutting Edge on Tuesday night , further allegations of fraud and deception emerged against the miracles industry in South Africa ,
20    or imprisoned in the investigation is a tale of illegality and deception on such a grand scale it makes the 1972 Watergate scandal ,
21        Birds , Rebecca ) have always been laced with intrigue and deception , but none quite like My Cousin Rachel . It 's a
22         , but also suggests that Nigerians have been fed lies and deception over the years by government and the NNPC officials . It also
23     even accuses Hillary Clinton and her campaign team of lies and deception . # " What I did was at the direction of and
24             in having fallen prey to a movement based on lies and deception . The wider question is about what kind of world we want
25                      support of the people . # " They use lies and deception as their crutch , casting me as some type of one-man political
26                   get the support of the people . They use lies and deception as their crutch , casting me as some type of one-man political
27          that Nigerians have seen through the falsehood , lies and deception that have characterized his incompetent and corrupt
28                    it . And thus the mayhem in society . Lies and deception in time will bite . Cheers Welcome to the discussion .
29                 with the facts , " Letourneau said . " Lying and deception by a police officer is an extremely unsettling character trait .
30   Mr Fong , whose client faces charges of market manipulation and deception , placed the spotlight on the flow of information in the market
31   Mr Fong , whose client faces charges of market manipulation and deception , placed a spotlight on the flow of information in the market
32                  good " ? Is this yet more smoke and mirrors and deception ? # Alan Greenspan recently said that he now thinks all this
33     # Thoughts , hidden , are the provenance of misjudgement and deception . Many are misjudged because no one sees their thoughts . Also
34       book titles include Hidden Sins , Secret Lies , Reckless and Deception . # Ms Abrams -- who received high-profile endorsements from
35     . Of course , this posting allows for false representation and deception . # Through the Internet or use of mobile phones , falsehood
36         repeated letdowns , erosion of authority of the state and deception . # Akhtar said that Jaitley should know more than anybody else
37                   2016 . # Asked if she needs to add strokes and deception to dominate world badminton , Sindhu said : " Definitely yes.It
38     bureaucratic doublespeak or when that fails outright theft and deception have thwarted plans to protect the basin . It has got worse
39                 is soon plunged into a life of crime , theft and deception . # By finding success in the marketplace with her flawless
40       Tshisekedi presidential candidacy for fraud , trahison and deception . # To reach there the media needs immediately to speak loudly
41          giant 's failure to protect the privacy of its users and deception about who has access to user data coming out of last year
42   hotspots were Morayfield , where 163 complaints were made , and Deception Bay on 158 . REVEALED : Gympie animal cruelty hotspots ,
```

Fig. 1: Sample concordance lines of *deception* (co-occurring with *and* in the L1 position) in the NOW Corpus, retrieved April 2019.

(lines 30—31), *theft* (lines 38—39) and *trahison* (line 40). The lines suggest that deception is a negative concept. However, individual instances of *deception* can point to more specific cultural aspects that relate to current topics in public discourse.

An example of a more time-dependent instance is concordance line 41, which discusses deception tactics in relation to a company's *failure to protect the privacy of its users*. While this is a very specific type of deception, it also relates to wider discussions surrounding technological developments and the need for new privacy protection measures. Further privacy-related examples are found in concordance lines beyond the sample in Figure 1. Examples (1) and (2) mention *deception* in the wider context of data protection legislation, specifically mentioning GDPR, the EU's General Data Protection Regulation; similarly, POPIA in (2) refers to South Africa's Protection of Personal Information Act. The examples highlight particular types of deception (*machine deception*, *cyber deception*) that pose challenges for the technology and security sectors. These developments in deception strategies and the legislative reactions in the form of data protection laws simultaneously create opportunities for businesses – like the one mentioned in Example (2) – to specialise in protecting their clients from breaching the new regulations. Accordingly, these examples explain particular aspects of deception in relation to technology and legislation in 2019.

(1) "Machine **deception** does not just refer to machines deceiving humans, however. It also refers to machines deceiving machines – bots - and people deceiving machines - troll armies and click farms. Information propagation methods and click farms will continue to be used to fool ranking systems on content and retail platforms, and methods to detect and combat this will have to be developed as fast as new forms of machine deception are launched."

GDPR and building trust
The introduction of GDPR could make any such breaches extremely costly. The ICO is yet to issue a big fine for a GDPR breach, but David Francis, head of security at IT services provider KCOM, believes 2019 will be the year it happens for the first time. (NOW Corpus; Macaulay 2019)

(2) In addition to SIEM, Zenith Systems also deploys solutions for sensitive data management (for POPIA and GDPR, etc), cyber **deception** and cog-sec, which is designed to change, at a cognitive level, risky user behaviour pertaining to e-mail and Web use. (NOW Corpus; *ITWeb* 2019)

Example (3) illustrates another aspect of the discourse of deception, which frames deception as a wider cultural concept in relation to news media. Crucially, this example explicitly raises the question of terminology. Here we focus on the contribution that corpus linguistics can make to understanding deception as a cultural concept by analysing the discourse. Other corpus linguistic approaches focus on unpicking the linguistic realisation of deception strategies (cf. e.g. Archer & Lansley 2015; Popoola 2018).

(3) "By virtue of saying 'fake news,' we ask the question, 'Well, what is real news?' and you invite people to label everything they disapprove of as 'fake news,'" Hall Jamieson said earlier this year during an interview on CNN. "As a result, it's not a useful concept. What are we really concerned about? **Deception**. And **deception** of a certain sort that goes viral."

> The term "viral **deception**" can be abbreviated as VD, Hall Jamieson deadpanned before her audience in Steamboat Springs, Colo. And more than fake news, VD is something society doesn't want or need. But her play on words was also created as part of a deliberate strategy. (NOW corpus; Ross 2017)

In corpus linguistics, a distinction is sometimes made between 'co-text' and 'context': while the former describes the lexical environment surrounding the word or phrase under investigation, the latter refers to corpus-external "situational and cultural parameters involved in the interaction" (Tognini-Bonelli 2001: 87). However, these concepts cannot be neatly separated. Situational and cultural parameters in which a text is produced are not fully reproducible from a text, but are reflected in its lexico-grammatical patterns. Information about the source of the patterns such as the venue and time of publication is part of the text-external context that also contributes to the meaning-making. Examples (1) and (2) above originate from technological media outlets; (1) is from a website called the *ITWeb* (subtitled "Business Technology Media Company") and (2) from the online publication *ComputerWorld Hong Kong*. The NOW Corpus makes it possible to examine patterns from a wide range of publications beyond traditional news outlets. This wide scope of online data offers new opportunities for corpus linguistic approaches to discourse. At the same time, new digital data sources complicate the understanding of meaning in social contexts, because conventional boundaries between, for example, media and corporate platforms are increasingly blurred. In the above examples, we have focused on the text-internal context. The lexico-grammatical patterns around *deception* determine the meaning of the word. At the same time, they broaden out to reflect more complex social and cultural contexts and the fast-changing interactions that determine them. As the mentions of the GDPR data regulations above have illustrated, time is an important contextual factor. The semantic preference of "dishonesty" will likely hold for *deception* in a range of corpora, but the link to GDPR is a relatively recent one.

Our example of *deception* is what can be called a 'cultural key word'. Cultural key words express meanings that are important to a society. Williams's (1983) vocabulary of culture and society demonstrates how such words are recognisable even without corpus linguistics. But what corpus methods can do is base the analysis and description of their meanings on discursive evidence as found in corpora. A number of corpus linguistic studies have tackled examples of what might be termed cultural keywords, e.g. Bevitori (2010) on *climate change*, or Stubbs (2001) on *heritage* and *care*, among others. Cultural key words are examples that highlight particularly strikingly how co-text and context come together. But even if the focus is not on cultural key words, the underlying principle remains: the analysis of language that draws on corpora will always – explicitly or implicitly – reflect the intricate connection between language, society and culture. In their broadest sense, corpora provide cultural data. In the more narrow sense of individual studies, the notion of context will be more specific.

Three types of contexts

In this volume, we approach context from three perspectives: (i) discourse contexts and culture, (ii) lexical and grammatical contexts and (iii) learner contexts. While the discourse context takes the widest approach and so links to cultural concerns in the broadest sense, the phenomena captured through lexical and grammatical contexts do reflect cultural concerns by means of the corpora in which they are found. The learner contexts are a particularly good example of the links between co-textual and contextual observations, where the context is determined by the experience of more than one language and culture and the detailed analysis of lexical and grammatical phenomena requires a focus on very clearly defined co-texts, too.

In Part I, "Discourse contexts and culture", the "contexts" and "cultures" analysed range from seventeenth century England and the diachronic study of historical American English to various present-day settings in Europe, the US and the Caribbean. Beatrix Busse starts this section with her analysis of place-making patterns in Brooklyn, New York (Chapter 2). Her chapter takes an interdisciplinary approach that is grounded in corpus linguistics and urban studies. She analyses a corpus of interview responses on the basis of Busse and Warnke's (2015) 'urbanity model'. Like Busse, Dagmar Deuber and Eva Canan Hänsel (Chapter 3) also make use of interviews. However, unlike the approach taken in Chapter 2, their analysis utilises interview data from newspaper editors as contextualising information for the corpus linguistic analysis of Caribbean newspaper articles. The use of interviews shows how the context of text production is also relevant to the analysis of the texts that we find in a corpus. Chapter 4, by Alessandra Molino, studies corporate self-representation strategies in a corpus of sustainability reports. Molino's methodology adapts collocation and concordance analysis to analyse local textual functions, following Mahlberg (2007).

In Chapter 5, Helen Baker, Ian Gregory, Daniel Hartmann and Tony McEnery take a clearly interdisciplinary approach: the authors analyse the relationship between discourses and places in seventeenth century texts related to prostitution from a historical perspective using 'concordance geo-parsing'. Their approach complements Busse's. Busse considers place-making as an active process, where the discourse plays a crucial role, and Baker et al. relate insights from the study of discourse to the places where they are produced. The relational notion of discourse also becomes apparent in Wolfgang Teubert's chapter (Chapter 6). Teubert envisages an interface for a dynamically expanding corpus that allows all discourse participants (not just linguists) to observe paraphrases and intertextual links and thereby interpret ongoing meaning change. With such a tool, according to Teubert, corpus linguistics would make a clear contribution to the study of meaning. So Chapter 6 discusses more generally the merits of corpus linguistics for the study of meaning and the future of the field.

Part II shifts the focus from the wider discourse to more specific contexts and patterns in lexis and grammar. Gregory Garretson's chapter (Chapter 7) introduces the concept of 'family collocation', which is based on grouping 'families' of node words together for computing collocational measures, similarly to lemmas. The rationale for family collocation is twofold: on the one hand, it accounts for potential connections between related word forms in the mental lexicon and, on the other hand, it draws on the advantage that statistical tests on the families are more powerful than those on individual word forms. In Chapter 8, Hildegunn Dirdal analyses the translation of *ing*-clauses from English into Norwegian using the Multiple-translation Corpus. She finds that the translator's individual style is one of the factors contributing to the choice of Norwegian language structure.

Chapters 9, 10 and 11 all make use of the Corpus of Historical American English for diachronic comparisons. In Chapter 9, Leonie Wiemeyer investigates diachronic changes in the productivity of so-called ‘native combining forms’ – elements derived from lexical words used to form new words such as *-thon* in *marathon*. Wiemeyer finds that the productivity and frequency of combining forms is closely linked to the topicality of the concepts that they correspond to and can be caused by cultural factors, such as technological developments. Chapters 10 and 11 both study patterns throughout the whole COHA from the 1810s to 2010s. In Chapter 10, Mark Kaunisto and Juhani Rudanko focus on exceptions to ‘Bach’s Generalisation’ which says that object control structures require a noun phrase object. The authors explore the use of the covert object pattern with *advise against* in relation to Bach’s Generalisation and discuss pragmatic considerations. Magnus Levin’s chapter (Chapter 11) similarly traces the development of a particular construction in COHA, but his focus is on subjective progressives. Levin finds a lexical restriction in the increase of subjective progressives. While they do not increase overall, patterns of subjective progressives with the adverb *always* do rise in frequency. The chapter discusses both ‘language-internal’ (particularly colloquialisation) and ‘language-external’ factors (i.e. gender) of this development. Over the last few decades, corpus research has shown that lexis and grammar are not separate entities; this is also reflected in Levin’s findings.

Part III, “Learner contexts”, the final and largest section of the volume, presents different perspectives on the corpus linguistic analysis of learner English. The studies are situated in the national contexts of various countries: in particular Norway, but also Germany, Japan and Sweden. Several chapters directly compare patterns between learner and native speaker writing. Chapters 12 and 13 both expand on Hasselgren’s (1994) notion of ‘lexical teddy bears’ – specific words that learners cling to even when these are not the best choice. In Chapter 12, Tove Larsson maps out the syntactic types of the ‘introductory *it* pattern’ in native and non-native student writing, finding that learners tend to (over)rely on one syntactic type of the pattern (“Subject + Verb + Complement”). Larsson argues that because this type can be considered a ‘lexico-grammatical teddy bear’, more emphasis should be placed on teaching the variety of types and the associated discipline-specific conventions. Chapter 13 by Hilde Hasselgård is also concerned with a comparison of patterns between native and non-native student writing, but with a focus on four-word lexical bundles. Hasselgård’s ‘phraseological teddy bears’ are therefore lexical bundles that learners use more often and in more contexts than native speakers.

Rolf Kreyer’s chapter (Chapter 14) takes an innovative approach to investigating learner texts by analysing conceptual and formal learner revisions in the Marburg corpus of Intermediate Learner English (MILE). Kreyer emphasises that the analysis of revisions provides insights into the writing process and the learners' difficulties, which are masked when only the final product is considered. The corpus design of the MILE is central to this approach to revision analysis. Text-internal markup that specifies alterations made during the writing process provides a new source of data on learners' interlanguage. Additionally, the longitudinal data collection makes it possible to study the development of learners' revision forms as they move to higher grades in secondary school.

In Chapter 15, Sylvi Rørvik examines the use of 'marked themes' – a resource for creating texture in systemic-functional grammar – in the writing of advanced Norwegian learners of English. The chapter applies the procedure of Granger's (1996) 'Integrated Contrastive Model' in order to compare argumentative texts from the Norwegian component of the International Corpus of Learner English (ICLE) with English and Norwegian expert and novice L1 writing. Chapter 16, by Susan Nacey and Anne-Line Graedler, also examines the Norwegian component of the ICLE, in addition to the Louvain International Database of Spoken English Interlanguage (LINDSEI). In their chapter, Nacey and Graedler compare the use of phrasal verbs by Norwegian learners of English across spoken and written modes. The results challenge the general perception that phrasal verbs are highly problematic for learners and suggest that the advanced Norwegian learners successfully use conventional metaphorical extensions of particle meanings.

In Chapter 17, Keiko Tsuchiya similarly studies spoken learner English, but places particular emphasis on a systematic corpus linguistic approach to studying multimodal corpus data. Tsuchiya's small-scale 'conversational gesture corpus' contains time-aligned transcripts of learner conversations and has been annotated with speakers' gestures and repair strategies. Multimodality is a growing area in corpus linguistics. While itself focusing on specific conversations, the methodology of Chapter 17 suggests connections with more large-scale applications in the space of corpus-assisted discourse analysis (cf. Bednarek & Caple 2017). Ute Römer rounds off Part III – and the volume – with a call for mixing methods in the area of Second Language Acquisition (SLA) (Chapter 18). Her chapter synthesises the approaches and results of two case studies in SLA research in which corpus linguistic approaches were combined with work from psycholinguistics, computational linguistics, genre analysis and cognitive linguistics.

Overall, the chapters in this volume illustrate a range of contexts to be explored by corpus studies. Putting discourse, lexical and grammatical as well as learner contexts together does not only highlight the breadth of the discipline but also encourages us to consider the connections. Corpora are much more than datasets. They are the basis for a contextualised approach to language.

References

Archer, D., & Lansley, C. (2015). Public appeals, news interviews and crocodile tears: An argument for multi-channel analysis. *Corpora*, *10*(2), 231–258.

Bednarek, M., & Caple, H. (2017). *The Discourse of News Values: How News Organizations Create Newsworthiness*. Oxford, UK: Oxford University Press.

Bevitori, C. (2010). *Representations of Climate Change: News and Opinion Discourse in UK and US Quality Press: A Corpus-Assisted Discourse Study*. Bologna, Italy: Bononia University Press.

Busse, B, & I. H. Warnke. (2015). Sprache im urbanen Raum. In E. Felder & A. Gardt (Eds.), *Handbuch Sprache und Wissen*. (pp. 519–538). Berlin, Germany: de Gruyter.

Davies, M. (2013). Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. Retrieved from https://www.english-corpora.org/now/ (last accessed May 2019).

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies* (pp. 37–52). Lund, Sweden: Lund University Press.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258.

ITWeb. (2019, March 29). Zenith Systems launches LogRhythm NextGen SIEM Platform at ITWeb Security Summit 2019. Retrieved from https://www.itweb.co.za/content/rW1xL755XEw7Rk6m (last accessed April 2019).

Macaulay, T. (2019, January 3). Data science and AI predictions for 2019. Retrieved from https://www.cw.com.hk/data-management/data-science-and-ai-predictions-for-2019 (last accessed April 2019).

Mahlberg, M. (2007). Lexical items in discourse: Identifying local textual functions of *sustainable development*. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora*. (pp. 191–218). London, UK: Continuum.

Popoola, O. (2018). Detecting fake Amazon book reviews using Rhetorical Structure Theory. In *Proceedings of MIS2: Misinformation and Misbehavior Mining on the Web*. Retrieved from http://snap.stanford.edu/mis2/files/MIS2_paper_20.pdf (last accessed April 2019).

Ross, T. (2017, July 24). Seminars At Steamboat presenter tackles fake news in an era of viral deception. Retrieved from https://www.kunc.org/post/seminars-steamboat-presenter-tackles-fake-news-era-viral-deception (last accessed April 2019).

Sinclair, J. (Ed.). (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London, UK: Collins.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, UK: Blackwell Publishers.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam, Netherlands: John Benjamins.

Williams, R. (1983). *Keywords: A Vocabulary of Culture and Society* (2nd ed.). London, UK: Fontana.

Part I: **Discourse contexts and cultures**

Beatrix Busse

Patterns of discursive urban place-making in Brooklyn, New York

Abstract: The aim of this paper is to conceptualise urban linguistic and semiotic patterns from an interdisciplinary perspective. Drawing on the example of multimodal discursive practices used in selected neighbourhoods of Brooklyn, New York, this paper shows how spaces are turned into meaningful places through various social semiotic moves and stylistic practices. These discursive processes of urban “place-making” (cf. Busse & Warnke 2015) create, construe and contest this specific urban Brooklynite place and identity and therewith mark – due to its speakers’ active role in positioning themselves in the social landscape (Silverstein 2003; Johnstone, Andrus, & Danielson 2006; Searle 1995) – the value of particular Brooklynite neighbourhoods. The paper chooses a mixed-methods approach which combines both a qualitative and a quantitative methodological framework as well as approaches from sociolinguistics, corpus linguistic methodologies as well as semiotic landscape studies.

1 Introduction

Figure 1 displays a tag that was photographed during my fieldwork trip to Brooklyn, New York, in May 2017. With rather unexciting fonts and a partially non-standard orthography it relays the sentence “Spread love, its the Brooklyn way”. Figure 2, also photographed on the same trip, is a photo from a mural of the semiotic landscape on N 8th St. and Bedford Avenue in the neighbourhood of Williamsburg, Brooklyn. It shows the same sentence but in a different artistic design. Despite the fact that both examples are different manifestations of that sentence, in each case, “Spread love, it’s the Brooklyn way” represents a patterned repetition of a linguistic strategy in the urban place of Brooklyn. This strategy is of value to and meaningful for both the producer and the viewer.

Beatrix Busse, Ruprecht-Karls-Universität Heidelberg, beatrix.busse@as.uni-heidelberg.de

https://doi.org/10.1515/9783110489071-002

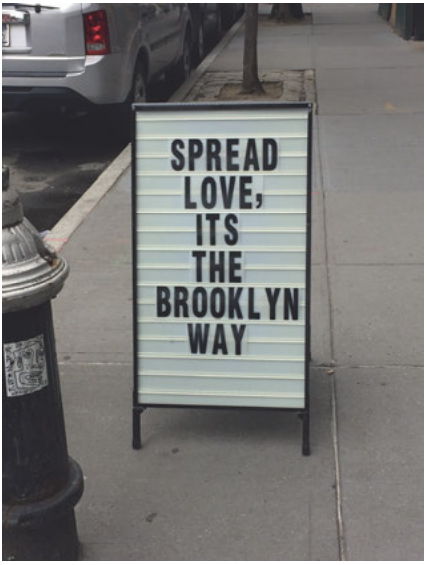


Fig. 1: Sign “Spread love, its the Brooklyn way”, Franklin Street ©Beatrix Busse.

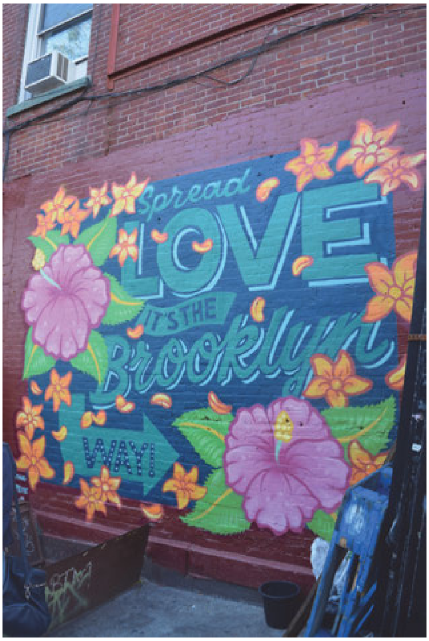


Fig. 2: “Spread love it’s the Brooklyn way”, N 8th St. and Bedford Avenue, ©Beatrix Busse.

The aim of this paper is to conceptualise urban linguistic and semiotic patterns from an interdisciplinary perspective and within both a qualitative and a quantitative methodological framework. I shall use the communicative complexities and semiotic patterns of the city, and specifically of the borough of Brooklyn, New York, as a testbed for this notion of patterns.

Drawing on Busse and Warnke's (2015) urbanity model, I propose that the emerging semiotic patterns of Brooklyn, such as "Spread love it's the Brooklyn way", act and are perceived as what I will call processes of "discursive urban place-making". These are linguistic and semiotic practices which turn a 'space' into a 'place' and make it meaningful for its inhabitants (cf. Cresswell 2004, 2006). Brooklyn is one of the largest, socially as well as ethnically most heterogeneous boroughs of the City of New York. Gentrification in Brooklyn has been particularly rapid in those neighbourhoods which, despite their geographical separation through the East River, are facing or are close(r) to Manhattan. My focus is therefore on exactly those now-gentrified neighbourhoods of Brooklyn, such as Williamsburg, Park Slope and Brooklyn Heights. I will show that various social semiotic moves and stylistic practices create, construe and contest this specific urban Brooklynite place and identity and therewith mark – due to its speakers' active role in positioning themselves in the social landscape (Silverstein 2003; Johnstone, Andrus, & Danielson 2006; Searle 1995) – the value of particular Brooklynite neighbourhoods. In other words, these patterned discursive strategies are *place-making*: They reflect and construe, (re-)define and (re-)evaluate Brooklyn and its gentrified neighbourhoods as a brand, "Brooklyn©", that is, a place with a particular character and style that has global impact on how other cities around the world perceive and create themselves.

Strongly linked to my notion of discursive urban place-making (through the use of linguistic patterns) is that of 'enregisterment' (Agha 2003, 2005; Johnstone 2009) which is the process by which linguistic features are ascribed with social values. In this paper, I will show that there is a need

1. to combine enregisterment with the concept of discursive urban place making,
2. to extend Johnstone's (2009) approach to enregisterment as pattern-based to all levels of language, not just to (local) dialect features or metapragmatic practices,
3. to demonstrate that enregisterment can actually be measured quantitatively with the help of corpus linguistic methodology and, at the same time, has pointedly singular qualities that interplay with its measurable features, and

4. that patterned generic reference to Brooklyn as a whole and in comparison with Manhattan mark strategies of discursive urban place-making which create Brooklyn© as a brand.

On the one hand, this study establishes norms and enregistered linguistic and semiotic patterned practices of discursive urban place-making. On the other hand, it is also situated in what Eckert (2012) calls the ‘third wave’ of variationist study and what Pennycook and Otsuji (Otsuji & Pennycook 2010: 246; Pennycook & Otsuji 2015) refer to as ‘metrolingual’ practices: the inclusion of singular contextual linguistic and semiotic artefacts which act as or move towards becoming place-making forces because they adhere to already established patterns, deviate from them or function as emerging triggers of place-making in the sense of “indexical mutability” (Eckert 2012: 94) and functional semiotic styling.

My study also supports Britain’s (2017) criticism of the sociolinguistic “gaze” as being circular because, he argues, sociolinguistic analyses have been too much focused on linguistic elites. This bias, according to Britain (2017), coincides with the ideological view that the British English standard or norm pronunciation is “Received Pronunciation” spoken by exactly that elite. What I shall present in this paper is also a theoretical and methodological redirection and shift of the “urban sociolinguistic gaze” towards both urban linguistic singularity as well as plurality and variation which are complexly and heterogeneously embedded in pattern-like strategies of urban discursive place-making (see also Busse 2018) and can be made visible on different levels of analysis as well as in different types of linguistic and multimodal data. Hence, I propose an innovative methodology which combines quantitative and qualitative approaches to assess patterned discursive urban place-making and enregisterment processes and to show that these can indeed be measured. This methodology is the first to combine (i) variationist sociolinguistic approaches with (ii) (standard) corpus linguistic methodology, and (iii) semiotic landscape studies. I therefore include a variety of data collected in 2012 and ranging from a corpus of semi-structured interviews conducted with Brooklynites to literary texts and examples from the semiotic landscape of Brooklyn neighbourhoods. With the help of these data, I will address the following research questions:

1. What are the contemporary changing and stable linguistic and semiotic patterns that reflect Brooklyn ‘as a place’ in everyday contexts and that define and evaluate (i.e. enregister) – urban place in Brooklyn?
2. What are the multimodal means of ‘being Brooklynites’ and of creating a place and a sense of belonging? What discursive place-making activities can be observed? How do these index social value?

3. How and why do Brooklynites use processes of enregisterment to make a physical location a distinctive place (Busse & Warnke 2015; Cresswell 2004) that is phenomenologically dense with meaning, familiar and legible for its inhabitants, that styles authenticity and identity (cf. Eckert & Rickford 2001; Lacoste, Leimgruber, & Breyer 2014) and shows alignment with groups?
4. What are the (urban) values that are connected with (variable) language patterns?

In the next section, I will provide a brief outline of the theoretical context in which I situate the approach I propose in this paper. I will introduce the definition of 'patterns' that I will be working with, explain both the model of urbanity that I am drawing on as well as the notion of variational place-making, and show how all of these are linked to the concept of enregisterment. The next chapter will then present an application of the urbanity model by analysing repetitive practices of urban discursive place-making in a varied data set – interviews and examples from the semiotic landscape – in selected neighbourhoods of Brooklyn.

2 Urban Semiotic Patterns and the Urbanity Model

Patterns are conceptualised as temporally and spatially situated configurations of linguistic and other semiotic signs which are repeated in a similar way, which are perceived as repetitive, which undergo transformations in space and time and which have constructive and social potential. Corpus linguistic identification and analysis of linguistic patterns have to be seen in a complex analytical framework of repetition based on frequency and of the saliency of repeated structures in relation to established norms analysed in, for example, specific text types, varieties or speech communities, and established by measuring the statistical association of e.g. words to constructions (e.g. Stubbs 2002; Hoey 2005; Gries 2008; Ebeling & Oksefjell Ebeling 2013; Biber & Barbieri 2007; Biber et al. 1999). A collocation is a case in point because in combinations of more than one word, where the focus is on the interplay between lexis and grammar, frequency of occurrence may be one criterion, but not the only one, as patterns may be highly frequent but not necessarily salient. Also, repetition, occurring at least once, may carry foregrounded functions which need to be analysed. And, finally, linguistic and multimodal pattern analysis also depends on and is influenced by how speakers and hearers use and perceive these repeated structures.

Corpus linguistic research has therefore also increasingly focused on how linguistic and other semiotic patterns and their variation serve as a resource for construing social meaning and identity in context (e.g. Baker 2014, Brookes, Harvey, & Mullany 2016, Partington, Duguid, & Taylor 2013). In corpus stylistics (Busse 2013, Busse 2014, Mahlberg 2013), it is now possible to even measure style, stylistic patterns and patterns of foregrounding (that is, parallelism *and* deviation) of collocations or keywords, for example, on a much broader scale, while, at the same time, linking these both quantitatively and qualitatively to functions and meanings in (historical) contexts of, for example, characterisation or social styling (Coupland 2007). This corpus linguistic notion of the relationship between social meaning and measurable or salient repetitive linguistic structures is the crucial and innovative addition to analysing patterns of discursive urban variational place-making in urban spaces in general and Brooklyn in particular.

Discursive patterns serve as a resource in the construction of identity in life/place, they reflect local dynamics and are rooted in particular practices of place-making. As such, variation can be seen "as a reflection of social identities and categories to the linguistic practices in which speakers place themselves in the social landscape through stylistic practices" (Eckert 2012: 92). Busse and Warnke (2015) developed a model of urbanity to account for the specific characteristics of discourse in urban spaces and the constructive potential of (patterned) language for urban place-making. 'Urban' denotes a set of values of a (big) city and, as a concept, it describes an evaluative relation rather than being an objective, descriptive term. Following Lefebvre (1974), Busse & Warnke (2015) work with three modes of urbanity that serve as interacting and interdependent parameters of urbanity. As shown in Table 1, these are (i) 'dimension', (ii) 'action', and (iii) 'representation'. 'Dimension' describes spatial dimensions in developed and open space. 'Action' is interpersonal and takes into account "lived experience, interaction and use of space by its inhabitants" (McIlvenny, Broth, & Haddington 2009: 1879). 'Representation' embraces the ways in which meanings are construed by means of cognitively represented and socially construed sign-making (Warnke 2013). These three modes are also connected and interdependent with the following six characteristic features: 'size', 'density', 'heterogeneity', 'simultaneity', 'multiformality' and 'multisemiosis'.

Tab. 1: Model of urbanity (Busse & Warnke 2015)

	Urban Modes	**Urban Characteristics**
Dimension:	size	density
Action:	heterogeneity	simultaneity
Representation:	multiformality	inter-semiotic

Let us return to the two photographs displayed in the introduction in order to explain the model. The photo of the sign in Figure 1 was taken on Franklin Street in the neighbourhood of Greenpoint. The sign was located in front of a coffee shop. Reference to the geographical coordinates marks the dimensional mode of the urban model. This information is of relevance because we know that it was found in a gentrified neighbourhood of Brooklyn. The actional and representational modes can be described as follows: the owner of this coffee shop saw the need to produce a sign in front of this coffee shop which contains – both in peculiar orthography and font – the combination of the directive (‘spread love’) and the assertive (‘its [*it’s* would be the grammatically correct choice] the Brooklyn way’), in which the generic reference to Brooklyn serves as a qualification of the noun *way* on this sign. What is the function of this utterance? It is a meaningful semiotic reproduction of a phrase used all around the semiotic landscape of Brooklyn and also used by the Brooklynite hip-hopper Notorious B.I.G. in his song “Juicy”. The phrase “Spread love it’s the Brooklyn way” has become not only popular all over Brooklyn but is also used to mark the Brooklyn identity on material objects. We have seen that it can be found on murals (Figure 2), and on many additional buildings, among others the so-called “Love building” in Clinton Hill in which B.I.G. was born. The urbanity model allows for the analytical as well as interpretational categories to analyse and describe how places, urban values and meanings are construed and reflected in the social interaction between urban space, materiality and what people do in it with language and other semiotic modes.

It comes as no surprise then, and we see it in the varying multimodal realisations of the “Spread love” examples, that urban spaces are characterised by *complexity* and *contrariness* (Venturi 1966). Urban spaces are multi-layered, heterogeneous and manifest a number of contrasts and divergences at the same time. In the “Spread love” examples, the linguistic pattern is the same – and we will see that even the generic reference to Brooklyn as a whole is a discursive pattern in Brooklyn, New York – but their material realisations are different: the number of different modes that interact in urban space, use of street art, and language.

As such, we, as analysts, are faced with a complex mass of different and heterogeneous types of data, ranging from language and official street signs to mass media, literature or street art.

These dimensions of urbanity as well as their accompanying characteristics also construe a high degree of linguistic and semiotic variation. Therefore, Busse and Warnke (2015) propose that the denser, more heterogeneous and inter-semiotic a city is, the more 'urban' it will be and the greater the (linguistic and semiotic) variation to be encountered. This relates to Vertovec's (2007) notion of the super-diverse city and the role of the English language therein:

> Two communicative domains in which English plays a dominant role seem particularly appropriate fields to put this recommendation into practice: the super-diverse global city (Sassen 1994) and the internet. The term *super-diversity* (Vertovec 2007) has been coined to mark the transition from patterns of multilingual and multicultural urban life that dominated until the last third of the 20th century to a more recent situation, in which ethnic and linguistic ties have become more diverse and also more unstable. (Mair 2013: 256)

The urbanity model is therefore not restricted to the city alone. Pennycook and Otsuji's (2015) approach to "metrolingualism" is also based on the assumption that "linguistic resources, everyday tasks and social space are intertwined" (Pennycook & Otsuji 2015: 14) in the process of creating identities of speakers and groups.

While places are, one the one hand, always historical and need to be studied within a historical continuum, they are, on the other hand, furthermore construed in their materiality as much as they simultaneously create it. Any source is materially visible or present as soon as it is said (cf. Blommaert 2013). Thus, (historical) urban linguistics as such becomes a part of the interdisciplinary field of *urban studies*. As urban space is 'linguistically loaded', urbanity has to be studied more extensively both from a synchronic and diachronic perspective.

Language and other semiotic modes do not create space, but rather place it in the context of the urbanity model. Both human geographers and proponents of urban studies differentiate between 'space' and 'place'.[1] A space can be defined as a context-free spatial dimension (cf. Busse & Warnke 2015). It refers to the structural and geometrical qualities of the physical environment. When human beings invest in a portion of space and become attached to it in some way, *space*

[1] Recent literature on place and place-making is vast (Friedmann 2010). Friedmann lists, for example, Jacobs (1962), Relph (1976), Heidegger (1993 [1977]), de Certeau (1984), Lefebvre (1974), Augé (2008 [1995]), Massey (2005).

becomes *place* (Cresswell 2004: 24). Thus, places are made subjectively and personally meaningful, they are constantly negotiated and embrace the (constructed) idea of 'home' (Cresswell 2004: 24, 39; see also Blunt & Dowling 2006). To charge an urban space linguistically is therefore an act of place-making as illustrated by the "Spread-love" examples.

Conceptualising the (social) styling of particular variational linguistic/ semiotic choices as a practice of urban place-making also prioritises the declarative function of language and discourse (Searle 1976) in urban space and its effect of shaping reality. The linguistic performative patterning in/of urban places is charged with a declarative force in the Searlean sense (Searle 1976), that is, it may comprise pragmatic constellations in which urban places are construed by means of linguistic and other semiotic sign-making. Not only urban protests, city guides, literary images of a city, songs, street art, toponyms, shop signs and so on can serve as examples of such a 'declarative city', but also how people talk about their neighbourhood or borough. According to Searle (1976: 13), a declarative not only creates equilibrium between the propositional content/illocutionary force of the utterance and reality (*word-to-world direction*), but also changes reality (*word-to-world- and world-to-word-direction of fit*):

> *Declarations*. It is the defining characteristic of this class that the successful performance of one of its members brings about the correspondence between the propositional content and reality, successful performance guarantees that the propositional content corresponds to the world. (Searle 1976: 13)

Stylistic practices in an urban place intend to be declarative in order to change reality, as they are part of the interacting modes of *dimension*, *action* and *representation*. Urban linguistics is therefore not necessarily interested in languages in the city, but rather in the relationship between language and urbanity.[2]

The relationship between language and space is reciprocal. Hence, urban spaces may act as parameters of variation, but language is also a parameter for the construal of urbanity by means of place-making practices. This embraces a concept of linguistic variability which places the social and constructivist function of language at its centre and also sees variation as "a social semiotic system" (Eckert 2012: 94), where linguistic and semiotic practices (ranging from, for example, standard to the vernacular, from literature to semiotic landscapes) of all sorts "are continually imbued with a variety of meanings" (Eckert 2012: 94) and are not "hierarchically nested within each other" (Kallen 2010: 42). They function

[2] See Busse and Warnke (2015: 520) for the distinction between 'urban' vs. 'urbanity'.

as stylistic practices and social styling whose meanings change, can be reinterpreted and recombined depending on either context or locality (see also Jaworski & Thurlow 2010: 1).

Thus, when human beings say things or make meaning through semiotic intervention, they create social relationships. Therefore, patterned urban styles of discourse dynamically create and may be associated with particular identifying characteristics of a city (Coupland 2007) and with (urban) place, even though this indexical potential and effect of language might not always be initially noticeable, consciously practised or even suggested directly. Agha's (2003: 231) conceptualisation of enregisterment as ideological processes whereby linguistic features become associated or linked with particular social categories (and can then be used to do social work (Johnstone 2009: 159)) is, in this way, a place-making process in and of itself. Hence, a set of repetitive linguistic strategies or profiles may function as "a linguistic repertoire differentiable within a language as a socially recognized register" which has come to index "speaker status linked to a specific scheme of cultural values" (Agha 2003: 231).[3] Therefore, if place is something that is made meaningful by human beings, enregisterment is a place-making activity.

The points of intersection between place-making and enregisterment are most obvious in the general correlation of spatiality with particular dialect features, which existing studies have highlighted:[4] Dialect linguistic features as such become enregistered (see e.g., Johnstone 2011, Beal 2009a, Zhang 2005). The other relationship foregrounds "metapragmatic practices" (Silverstein 1993) that show which particular linguistic features encode (local) identity and place. On the basis of a wealth of historical and contemporary linguistic data, Johnstone's (2009) work on "Pittsburghese" illustrates that a fixed one-to-one correlation between linguistic variation and demographic facts (cf. Johnstone 2014: 107) is not sufficient to explain when, why and how particular ways of speaking come to be

[3] Agha alludes to accommodation theory (Giles & Smith 1979) in which the fact that speakers modify their accents to either converge or diverge from an existing pattern is a major tenet. Le Page and Tabouret-Keller (1985) also address "acts of identity" invoked through language.

[4] Joyce (1991), for example, discusses urban dialects with iconic working-class figures, Simpson (1999) shows how pop music artists adopt a certain American accent for certain effects, and Beal (2009b) points out that the songs of the British pop group "Arctic Monkeys" are witty narratives of urban Sheffield life with a strong sense of place which can be seen in their pronunciation, their choice of lexis and reference to nostalgic places in Sheffield. In turn, Hall-Lew and Stephens (2012) demonstrate how, in contrast to the term *city*, *country* is used to enregister a number of language attitudes and American imaginings of particular/typical rural practices/personae and speech styles.

discursively construed as having a social value. Linguistic variation and difference can only be noticed when local people become more mobile, see the differences between their speech and that of others and begin to talk about said differences (Johnstone 2014: 107). This is then followed by a shift to new meanings and values associated with that particular set of features. Users construct a different evaluative subset and express local identity and projecting authenticity, nostalgia and local pride with it. In the process of enregisterment, the local linguistic sign – such as "Spread love it's the Brooklyn way" – is related to social meaning. This relationship depends on and is construed by the speech community. The fact that the linguistic sign co-occurs with the concept that it is taken to mean further enables the speech community to provide it with the potential to establish this indexical relationship. In the examples already discussed, "the Brooklyn way" to spread love, which is materially visible at several places in the neighbourhoods under investigation, sets itself apart as a Brooklynite (and, for example, not Manhattanite) urban practice of creating belonging and a sense of togetherness.

A "commodity situation" (Appadurai 1986) which is a "situation in which its exchangeability (past, present, or future) for some other thing is its socially relevant feature" (Appadurai 1986: 13–14) can also occur as a process of enregisterment. A linguistic variety or sets of varieties are available for purchase and people will pay for it. The "[l]arger cultural framework" as well as "the ideological and material contexts" (Johnstone 2009: 162) need to be outlined to show why it makes sense for people to buy and sell linguistic features. The "Spread love it's the Brooklyn way" example has not only been used as a musical commodity, but it is also a linguistic practice which has been materialised in the public urban place and in that it can be found on websites that offer graffiti and street art tours through Brooklyn neighbourhoods.

Specific values of a place can be created and enregistered through standard language and patterned repetition of linguistic structures and meanings on all linguistic levels interplaying with other semiotic modes. This indexical relationship has not yet been investigated pointedly as a process of enregisterment and as a means of describing urban place-making activities. One reason for this may be the variational linguistic bias and the focus on phonological and dialect features (cf. Johnstone 2010). Another reason can be seen in the neglect of using a corpus linguistic methodological framework for measuring (social) styles on all levels of language, which is what this approach adds to the analysis of patterns of discursive urban place-making. Answering the questions as to how and when discursive practices become enregistered so that speakers and agents in a material world perceive them as such and perform a certain identity with their use is a very complex and comparative enterprise in which variation and attitudes to it are prerequisites for the process of enregisterment to become recognised and adopted. In addition, not every patterned discursive practice indexes just one identity, and meanings may change over time as values may change: “repeated combination of stylistic complexes with socially located individuals and their activities and social moves establishes what seems a natural connection, leading to iconisation” (Eckert 2012: 92). Therefore, it is not a fixed and regular pattern of socio-economic stratification of linguistic forms that is the most crucial state to be assessed or analysed, but rather the observation that these are available for further indexical moves or more local categories. To capture “the social detail that vivifies language usage” (Moore 2012: 67) in general and enregisterment as variational place-making in particular, one cannot see the social styling of enregisterment as part of a fixed, hierarchical linguistic profile.

To conclude, patterned linguistic and multimodal discourses in the city have the potential for creating urban places that carry a diversity of social enregistered meanings. In contrast to earlier approaches, this paper proposes that practices of urban place-making can be investigated by a combination of quantitative and qualitative methodologies and that it is this interplay between sociolinguistic and corpus linguistic methods that provides us with more sophisticated tools to measure and describe the complexity and variety of urban data.

3 Method and Data

Focusing on (historical) variational urban place-making and accounting for the specific characteristics of the urban model demands a specific methodological framework. Such a framework has to embrace different levels of abstraction when it comes to the indexical potential of patterned discursive place-making practices and it has to be able to capture a continuum of singularity and iteration. In other words, it must embrace what Stubbs (2013: 30) describes as follows:

> The question is here how to relate Sinclair’s bottom-up empirical description of language use (Sinclair 1996, 1999) and Searle’s top-down analytic explanation of society (Searle 1995, 2010) in which he attempts to explain “the exact role of language” in the creation of social reality (Searle 2010: ix).

To find answers to this question is at the heart of my approach to discursive patterns of urban place-making.

One could consider any individual (discursive) act of place-making to be a declarative charging of space. At the same time, urban linguistics follows the general linguistic interest to analyse and interpret social representations and conventionalised rules. And yet, the urban linguistic focus is not only on repetitive quantifiable and therefore measureable patterns of linguistic usage which can be seen as formally marked agreement of usage visible in iterative linguistic practices in, for example, mass media such as newspaper discourse, and analysable with the help of quantitative and corpus-assisted investigations. The focus is also on formally singular expressions, which can be traced back qualitatively – and on different levels of abstractions – to patterned place-making practices. This does not entail a contrast in urban linguistics between quantitative corpus-based investigations of patterned linguistic constructions and qualitative ethnographic work, for example. On the contrary, in order to account for the parameters of complexity and contrast in the urban model, analytically speaking, it is necessary to collect a variety of data and to include quantitative and qualitative methods in a balanced way. Methodologically, this is then represented by a triangulation of data. Part of a conceptualisation of discursive patterning and place-making is the level of abstraction of a patterned feature, unit, concept or process. It also has to address the fact that frequency is relative, and comparative. Therefore assessing, measuring and analysing discursive patterning and its variation and change in urban space should be both quantitative and qualitative, interdisciplinary in approach and methodology, and above all contextual. Discursive patterns, their variation and change in urban space are social, scalar, contextual, aesthetic, indexical, and mobile.

Hence, in order to find answers to the research questions outlined at the beginning I draw on classic methodology from sociolinguistics, using semi-structured interviews with Brooklynites, and combine these with methods from semiotic landscape studies, using exemplary photographs I took on selected streets of Brooklyn which focus on how Brooklyn as a place is conceptionalised semiotically. Classic corpus linguistic methodology is employed by applying keyword analysis to the interviews. The idea is to ask whether discursive urban place-making strategies can be quantitatively measured by looking at the words in the interviews that are statistically speaking more frequently occurring when compared with another corpus of a similar genre. The assumption behind this methodological focus is that a statistically marked repetition of particular lexical items may point to language practices that are enregistered and act as discursive urban place-making practices in urban space. The results of the quantitative analyses are then levelled against more singular practices of urban place-making in the semiotic landscape in order to test whether despite their singularity, they still display similarly patterned features of discursive place-making on different levels of abstraction.

Let us look at the data in more detail. It is impossible to investigate place-making strategies in all of Brooklyn's neighbourhoods. Brooklyn is socially heterogeneous and New York's largest borough with 2.5 million inhabitants. While, for example, the neighbourhoods closer to Manhattan, such as Park Slope, Brooklyn Heights or Williamsburg, were gentrified at different stages in the second half of the 20th century and middle-class intellectuals or hipsters have flooded these neighbourhoods (cf. Franz 2015; Osman 2012), gentrification has recently happened or has yet to take place when moving further into the south of Brooklyn, to Bushwick, for example. Also, it should be stressed again that – in the dimensional mode – Brooklyn's geographical location on the map of greater New York is an island, separated from Manhattan through the East River but connected to it by the Brooklyn Bridge or the Williamsburg Bridge. There is also the frequently criticised L-Train and recently, as another result of gentrification, a free water taxi ride from Manhattan to Brooklyn sponsored by IKEA. Hence, a certain amount of mobility is needed in order to leave one neighbourhood for the next – an aspect of separation that is not only a geographical fact but also mentally present in the minds of the Brooklynites, as will be shown below.

The examples presented here are taken from the gentrified northern Brooklyn neighbourhoods Williamsburg, Park Slope and Brooklyn Heights. Drawing on these data, I have not yet been too adventurous in the sense that I have been focusing on neighbourhoods closer to Manhattan, which, as we will see below, has an effect on the ways the neighbourhoods are construed as places. The interview data was collected on an earlier fieldwork trip to Brooklyn in February 2012. Further fieldwork has been conducted since. I set out to interview Brooklynites in the streets of the different Brooklyn neighbourhoods mentioned above. The social demographic of the 58 white interviewees turned out to be biased towards artists, designers, writers, and dog walkers mostly between the age of 20–40 and amounted to a small, yet revealing corpus of 33117 words - the Brooklyn Corpus (2012). The questions I asked in the semi-structured interviews were:

1. Do you live in this area?
2. How would you describe this area?
3. Why did you move to this area?

In order to identify keywords, this corpus is compared to a spoken sub-corpus of the American National Corpus.[5] The collection of data also focused on qualitative

[5] Reference corpus data is taken from roughly the same period, the same genre and ideally the same variety of English. The American National Corpus (2005; http://www.anc.org/) contains

sampling of the semiotic landscape (Jaworski & Thurlow 2010) in those neighbourhoods, as a first systematic step of my analysis of place-making in Brooklyn. In Kallen's (2010) terms, I mainly compiled individual semiotic data from the 'civic frame', that is, from public street signs, from the 'marketplace frame', which refers to signs from commercial enterprises, and from the 'wall frame', in which individual and disparate stickers, artwork, graffiti and temporary posters are considered the "prime mode of expression" (Kallen 2010: 43). This paper focuses on street signs.

4 Discursive urban place-making: Enregistering the Brooklyn© brand

In this section, the aim is threefold:

1. I will show that discursive urban place-making strategies, which have become enregistered, can be detected and measured with the help of corpus methodology, e.g. keyword analysis.
2. Going beyond Johnstone (2009), I will show that enregisterment must not necessarily be linguistically realised by means of phonological features or those representative of a local dialect, but can be practised repetitively on various levels of language.
3. On the basis of a keyword analysis of the Brooklyn Corpus (2012) I will show two particular discursive urban place-making strategies which turn Brooklyn© into a brand: one is the generic reference to Brooklyn as a whole, the other is the comparison between Brooklyn and Manhattan. Both practices which are quantitatively identified to be patterned achieve further prominence by my qualitative analysis of selected samples from the semiotic landscape of Brooklyn neighbourhoods.

Table 2 shows the results of the keyword analysis of the interview corpus compared to a spoken sub-corpus of the American National Corpus.

the spoken Charlotte Corpus of face-to-face conversation representative of Mecklenburg County and North Carolina compiled in the 1990s. It contains roughly 200,000 words.

Tab. 2: Keywords in the Brooklyn Corpus (2012) (reference corpus Spoken Charlotte Corpus of the ANC).

	keyword	absolute figures	keyness value
1	*like*	667	719,09
2	*Brooklyn*	167	717,62
3	*it's*	394	585,80
4	*Manhattan*	88	351,28
5	*here*	195	319,22
6	*more*	153	260,24

One could object to the significance or foregroundedness of these purely numerical findings by arguing that there is no quantitative 1:1 correlation between quantity and foregrounded meaning and that the reference corpus chosen is not the ideal source, especially as it is a challenge to find an appropriate spoken reference corpus in terms of size, register, genres and variety of English.[6] Also, if interviewees are asked to describe a location, one could argue that reference to its name as a geographical location is the obvious choice and so Brooklyn would be an unsurprising keyword. However, note that first interview question asks “Do you live in this area?”, i.e. not “Do you live in Brooklyn?”. Interviewees do have other terms at their disposal when referring to the neighbourhood (and different neighbourhoods promote their locality linguistically by referring to the neighbourhood they belong to). Possible alternatives include: New York, the name of the neighbourhood they are in, but also more specific terms such as “this block” or “this neighborhood”. My interviewees frequently and repetitively choose the proper name *Brooklyn* to refer to it as a concept with an accompanying identity. This practice involves meaning-making because the name *Brooklyn* is generically referred to and can be seen as a construction of place deixis, if not person deixis, to associate textual scope with spatial and biographical identities and to accumulate symbolic as well as cultural capital. Brooklyn is given a voice and this choice reflects how Brooklynites think they should talk about it. One interviewee, a 32-year old male waiter and dancer, even construed Brooklyn© as a brand:

[6] I am in contact with Mark Davies to discuss means of uploading my Brooklyn Corpus to the Corpus of Contemporary American English.

(1) I think like that is why Brooklyn is getting so much press, and it is like it is and then like with Brooklyn industries the clothing brand that is like there is more of an awareness of Brooklyn now and and it is almost become a brand. (24)

A specific grammatical strategy, which further realises the branding of Brooklyn, accompanies the generic reference to Brooklyn in the interviews. Often, *Brooklyn* is used as the attribute, the carrier or an adjunct of place in relational clauses. Examples are: *Brooklyn is*, *There is…* [in Brooklyn], *Brooklyn has* or it is referred to in the clitic *it's*. These are relational clauses (Halliday & Matthiessen 2004: 211). However, relational clauses unfold a process of being in which something is said to be something else. The relationship is (statically) established between two different entities and experiential weight is set up in the participants. Identifying relational clauses identify the carrier further and in attributive intensive relational clauses the two elements often differ in generality but they are seen to be on the same level of abstraction. The following dialogue, which I conducted with a 33-year old female designer, illustrates the generic usage of *Brooklyn*. At the same time, the reference to Brooklyn as being "hot", which, according to the OED may also refer to something being sexually attractive (OED 2018: II.8.c), shows the extent to which it is seen as an attractive borough to live in. Brooklyn is personified. It is acting and attracting and relates to a human experience.

(2) "Brooklyn now is having a big revival. It's very hot."
"In what sense revival? "
"Um, artistically, the restaurants, a lot of young people flocking to Brooklyn now. I think there is a very exciting vibe in Brooklyn now." (26)

Here is another example from a 32-year old female designer, which additionally abounds in a number of hesitation fillers (cf. Biber et al. 1999; Busse 2010), such as *you know* or *like* that may indicate insecurity or be a typical marker of American English spoken discourse:

(3) Like a lot of like it's very like homegrown, like grassroots, like a lot of people are like you know, if it's like you know oh this isn't made in Brooklyn or handmade in Brooklyn, it's like a cool […] a little bit of an authentic city. (78)

I will only focus on this particular use of *cool*, which refers to "attractively shrewd or clever; sophisticated, stylish, classy; fashionable, up to date; sexually attractive" (OED 2018: 8). *Cool* also collocates with authenticity in the relational clause.

The keyness of *cool* (29, keyness 46.14) in the Brooklyn Corpus (2012) supports Zukin's (2010) observations on "How Brooklyn became cool". According to Zukin (2010), Brooklyn developed from a "gritty" to a "cool" place, from a place where people come from to a place where people come to because it has to be understood as a place of creative consumption:

> I think it was the consumption spaces, the stores, bars and cafés where you could look through plate-glass windows and see people living a kind of aspirational life, but in a low-key affordable way. Brooklyn became to be understood as a place of creative consumption. (Zukin 2010: 35)

In the interview corpus, the enregisterment of "Brooklyn-as-a-commodity-name," is created by means of a continuous mentioning of the Brooklyn© brand in line with or contrast to Manhattan. This is done by a number of repetitive patterns, including comparisons as well as the use of spatial and personal deixis and negative polarity. In comparisons between Brooklyn and Manhattan, a frame of reference (Halliday & Matthiessen 2004: 560) is provided, either by features of identity, similarity and difference or by particular features of quantity and quality. In the interviews, *Manhattan* is also among the key lexemes (88) with a keyness value of 351.28 (see Table 2), as is the periphrastic comparative particle *more* (158; with a slightly lower keyness value of 260.24). Furthermore, the qualities that generally distinguish Brooklyn from Manhattan are economic and life style criteria. Examples are "[i]t's like much more affordable Brooklyn still and you got a lot more space" (78) given by the already mentioned 32-year old female designer, where the adjective from the semantic field of economics "affordable" refers to the lower costs of living compared to Manhattan and "a lot more space", dimensionally speaking, refers to larger housing facilities. The statement "I just like the life style better in Brooklyn. I feel like you get more for your money, it is a little bit more laid-back" (19), which was expressed by a 34-year old male studying to become a pre-school teacher, argues along similar lines.

Subjective feelings and criteria of what constitutes home and a sense of community are also repeatedly mentioned in a declarative vein. These can be attributed to what I will call a discourse of counterurbanisation in which an 'anthropological place' in Augé's (Augé & Colleyn 2006) sense (rather than a 'non-place') is created through constant repetition of the more home-like character of Brooklyn. This was even expressed by a male interviewee who works in finance: "Actually, probably, it's a completely different atmosphere. I think it's more … less citified than Manhattan." (30). Two further examples, expressed by a 27-year old male media expert and a 30-year old male brand director at a creative agency, are:

(4) Because you get all – like there's all the activities that you want to do without the bullshit, like there's a little – you are less inundated by crowds, a tiny bit less inundated by advertisements, you know, there's a pace here that's like still busy and competitive, but not just like, not so consistent at 24 hours of the day, you know. (24)

(5) Manhattan is gross. Brooklyn is just like, it's more like community, it's more like neighborhoody. Ahm, you can turn off here in Brooklyn, sometimes. You always have to be like on and working and networking when you are in Manhattan. So it's a little bit more relaxed, we just like the vibe better and we like the neighborhoody aspect of it. (36)

These verbal creations of place also allude to the 1940s when authors, for example, considered Brooklyn to be an alternative, cheaper, less competitive, less crowded, less touristy place where journalists and bohemians could find a home (Zukin 2010: 40).

In the established comparisons outlined so far, Brooklynites both talk about what is in Brooklyn and what is not in Manhattan or what is found more in Brooklyn and what is found less in Manhattan. Brooklynites seem to have a "particular reason for talking about what is not rather than what is" (Thompson 1996: 56) – a phenomenon that is strengthened in the use of negative polarity, where Brooklyn is defined as not being like Manhattan: "I wouldn't wanna live in Manhattan, it's like that space is, there is no green, there is no parks" (72). The effect of the comparative and contrasting structure through the syntactic negation *would not* and a negative particle *no* is interpersonal/attitudinal (Givón 1993: 202). Psycholinguistic research (Clark & Clark 1977: 110) and cognitive linguistics (Lakoff 2004, 2006) have illustrated that negatives (and perhaps even contrasts) also entail the positive counterpart, that is, "when we negate the frame, we evoke the frame" (Lakoff 2004: 3). In other words, in this example, the negation of the neighbourhood frame for Manhattan evokes it for Brooklyn, which is further reinforced through the mental process *wanna*. The "special purpose" (Leech 1983: 101) of the use of these negatives is to let Brooklyn shine. Jordan (1998: 707) states that

> an understanding of the meaning of negation [...] can no longer be separated from their contextual and interpersonal functions in real language. Especially for the use of negation in texts greater than a sentence or two, the contextual and textual roles of negation become of paramount importance. [...] This involves an analysis of how negation is used in definable patterns of communication in various forms of actual written and spoken English - from single statements to inter-document connections of meaning.

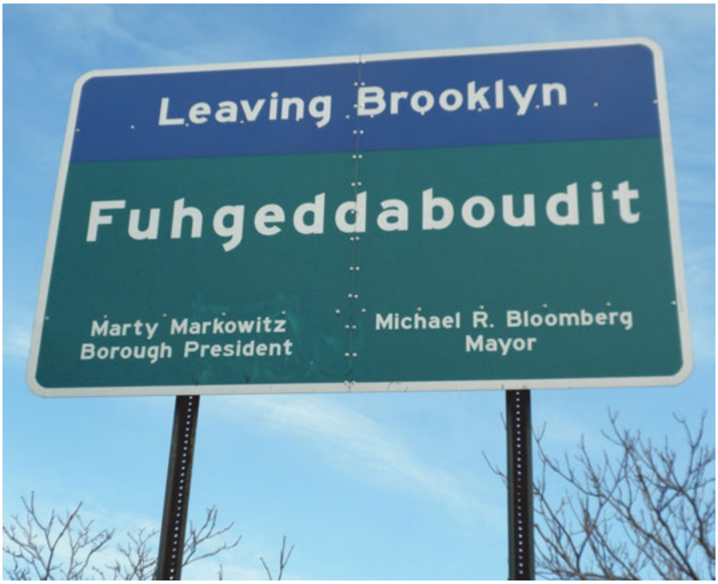


Fig. 3: “Leaving Brooklyn – Fuhgeddaboudit” – Road Sign on Belt Parkway, before crossing Old Mill Creek Bridge into Queens, Burger International Photography.

The contrast to Manhattan “haunts the island like a doppelgaenger or a conscience” (Capote 2001: xvi). The narrator Nathan in Auster’s (2006) *The Brooklyn Follies* also explains one character’s decision to move to Brooklyn by drawing on the contrast: “[h]e chose Brooklyn because it was New York and yet not New York” (Auster 2006: 50).

To verify and enhance the quantitative findings of the keyword analysis, I will qualitatively analyse two material artefacts from the urban semiotic landscape of the neighbourhoods under investigation. These examples illustrated in Figure 3 and 4 lend further support to the identification of an enregistering discursive urban practice which echoes the linguistic strategies outlined and quantified above and which is not just an example of the use of dialect features or of nostalgic references to the past. The examples in Figure 5 and 6 are taken from what Kallen (2010) calls the ‘civic frame’, which marks the official state-regulatory sign-making in terms of “labeling and delimiting territory and in regulating behavior” (2010: 43). In Brooklyn, the purpose of these street signs is not to organise and regulate traffic. Instead, the two examples display a declarative intervention that functions as enregisterment and helps build up and construe the Brooklyn© brand to thousands of mobile city dwellers who come to and leave

Brooklyn every day: road signs at Brooklyn motorway exits and entrances enregister Brooklyn through a generic reference to Brooklyn as a whole and quoting a number of fixed and authentically historical Brooklyn phrases such as “Fuhgeddaboudit” and “How sweet it is”.

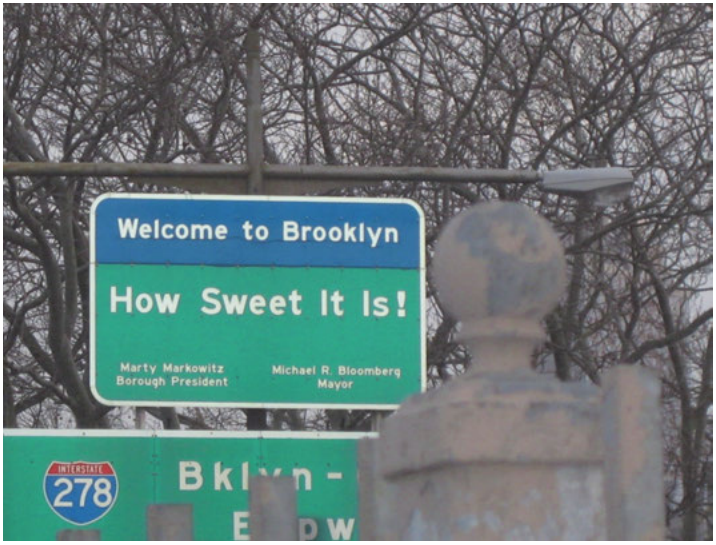


Fig. 4: “Welcome to Brooklyn How Sweet it is” – Road Sign Bridge on Brooklyn Bridge, southbound exit, Cadman Plaza West, Flatbush Avenue, © Beatrix Busse.

“Leaving Brooklyn? Fuhgeddaboudit” is an urban hip colloquial dialectal expression which means something like “the issue is not worth the time, you better stay!”, the world beyond Brooklyn, which also includes Manhattan, is not as good as Brooklyn and thus not worth bothering with. Written so as to represent the famous Brooklyn accent, this sign purposefully disregards orthographical and phonological standards (e.g. of General American). The narrator Nathan in Paul Auster’s (2006: 12) novel *The Brooklyn Follies*, for example, pointedly highlights that there are also negative perceptions of the local Brooklyn lingo – “that unmistakable accent so ridiculed in other parts of the country, which I find the most welcoming, most humane American voices.” However, while the Brooklyn accent is made fun of outside of the borough, it is celebrated within it and has achieved a status of enregisterment marking identity and place. The narrator in Auster’s novel clearly subscribes to this view.

A historically motivated, but still authentically Brooklyn strategy of semiotic place-making and enregisterment is also illustrated by the road sign "Welcome to Brooklyn - How sweet it is!", found on the Brooklyn Bridge, southbound exit, Cadman Plaza West, Flatbush Avenue. "How sweet it is" was the catchphrase of Brooklyn entertainer and television star Jackie Gleason (1916–1987). His show *Honeymooners*, which began in 1952 with sketches he revived until 1978, was set in Bensonhurst, Brooklyn. This quote within the 'civic frame' (Kallen 2010) on a highly public road sign declares the Brooklyn© brand by a nostalgic reference to the past. It reinforces a sense of local identity, illustrates the success of Brooklynites in the media, and at the same time stresses Brooklyn's historical past and tradition. Discourse about those road signs reinforces their potential for enregisterment. Borough President Markowitz commented on the branding function of these road signs:

(6) Once you enter Brooklyn, there's no good reason why you should ever leave. [...] These signs are just another great example of the Brooklyn attitude, and they capture the spirit, energy and enthusiasm alive and well all across Brooklyn. It also gives people one last chance to turn their cars around and stay in the promised land.[7]

In these two specific examples, we have seen that generic reference to the borough as a whole is a place-making strategy which is repeated and has pattern-like status. Here it is paired with authentically Brooklyn phrases which reinforce and enregister Brooklyn identity and the brand. Enregisterment is a communicative and social practice because a number of confluent practices merge and emerge (Carter 2012; Coupland 2007; Moore 2012). Therefore, processes of enregisterment (assessing, analysing, measuring and interpreting) are in need of a methodological framework which enables these kinds of analyses. It is crucial to take account of the 'communities of practice' (Wenger 1998) (rather than the speech community) in which language usage, the use of particular styles, or the use of genres take place. Coupland (2007) and Moore (2012) point out that only in a specific community of practice can linguistic features become socially meaningful (Moore 2012: 71). This entails understanding the social concerns of a historical community, how they are embodied in historical social styles (Moore 2012: 71) and which linguistic features occur in interaction with others (Moore 2012:

[7] See also Popik (2006) for other road signs.

68). In other words, in Brooklyn, I have investigated practices of how people describe a particular place and what they do with language to create that place. While it is always possible that speakers – past and present – exhibit particular stylistic effects or characteristics outside their socio-economic classification, the social meaning of a linguistic feature (and a genre) is typically underspecified until it enters into a speaker's or a group's social practice (Moore 2012: 68). Thus, the generic reference to *Brooklyn* and its comparison with Manhattan has entered this linguistic practice of place-making by a group of Brooklynites.

In this respect, more institutionalised contexts and everyday language usage and local practices have to be seen on a continuum of discursive urban place-making. Otsuji and Pennycook (2010: 3) have derived the concept of metrolingualism from "metroethnicity" (Maher 2005) to refer to "creative linguistic conditions across space and borders of culture, history and politics, as a way to move beyond such terms as multilingualism and multiculturalism". This claim is of crucial importance both in relation to the theoretical framework proposed here as well as to the practical analysis of patterned discursive urban place-making strategies. Heterogeneity characterises urban space, because people of different and mixed background use, play with and negotiate identities through language (Otsuji & Pennycook 2010: 244). As such, it is necessary to explore "how such relations are produced, resisted, defied, rearranged; its [metrolingualism's] focus is not on language systems but on languages emergent from contexts of interaction" (Otsuji & Pennycook 2010: 246). Since this multilingual and multi-semiotic sphere centres on the everyday use of mobile linguistic practice where patterns are adapted as forms of social styling that produce linguistic repertoires (Li 2011), it embraces the relationship between linguistic practices, everyday tasks and social place.

5 Conclusions

As a process of discursive urban place-making, enregisterment is construed through various semiotic modes and patterned stylistic practices. In the Brooklyn neighbourhoods investigated in this study, branding linguistic strategies go beyond using dialect and repeatedly include generic references to *Brooklyn*©, identifying clauses with *Brooklyn* functioning as the carrier or the identified as well as comparisons between Brooklyn and Manhattan, which, on the lexico-grammatical level, interplay with reference to semantic domains of community life-style, arts and the past.

This chapter has shown that investigations of the process of enregisterment cannot be restricted to the analysis of dialect features or metapragmatic practices alone, but – in order to break with the traditional variationist sociolinguistic bias of the first wave (Eckert 2012) – must be broadened to both the analysis of (standard) patterned linguistic constructions on all levels of language and a combination of sociolinguistic and corpus linguistic methodologies. These are crucial for assessing, analysing and disclosing new and other patterns of urban discursive place-making, which would otherwise have gone unnoticed. The interplay between the quantitative and the qualitative has been positioned as both a new methodological framework and meaning-making processes of discursive urban place-making. It is this very interplay that can be made visible on various levels of abstraction and which in comparative exercises oscillates between singularity and iteration. As such, Eckert's concept of "indexical mutability" (Eckert 2012: 94) of patterned variables is inherent to discursive place-making, as these are not markers of fixed, but rather of mutable meanings, which change in contexts of production and reception.

Selected interviews with inhabitants from the neighbourhood of Bushwick further underline this need for a change of gaze. Speakers have, by means of the linguistic practice of comparison, revealed and construed familiar discursive patterns of enregistering Bushwick as a valuable franchise of the Brooklyn© brand. And yet, there seem to be fewer generic references to Brooklyn as a whole and the basis of comparison appears to have changed from Manhattan to Williamsburg, which – at least in those parts facing Manhattan – is no longer the cheap and unexciting working-class and immigrant neighbourhood it once was, but highly gentrified.[8] Rich 'kids' are attracted to Brooklyn because of its proximity to Manhattan and because expensive condominiums have been built. On the famous Bedford Avenue, you find boutiques, art galleries, restaurants, whole food shops etc. Bushwick is thus seen as the "next Williamsburg". The change from a rundown Bushwick to a romantic, opportunistic, but up-and-coming-place that has yet to be discovered, is construed through lexical choices, verbs of movement and the present progressive to highlight that gentrification is happening, due to the

[8] It is also a neighbourhood which has a high Jewish population but the waterfront has been the target of urban city planners who have constructed a number of expensive condominiums to attract middle class bankers and other rich people. Gentrification can also be witnessed in parts of Bedford Street as well, which is characterised by art galleries, boutiques, warehouses, organic food shops, bars and cafés, but also caters to the needs of the middle class. Chains, such as *Duane Read* – a drug store – and *HSBC* have entered this area.

artistic scene and that money can still be made – depending on whether you are an outside or an inside viewer.

It is this aforementioned "indexical mutability" (Eckert 2012: 94) of linguistic variables, that is the transformation of values, that needs to be investigated further in a systematic way, as the use of the neighbourhood of Williamsburg functions as a new reference point for enregistering the value of neighbourhoods like Bushwick, whose newly acquired prestige is at least discursively construed. Both synchronically and diachronically – the idea of "moving into Brooklyn" and not being "pushed out" need to be the focus of future investigations of indexical place-making strategies. Also, systematic onsite visits to more central neighbourhoods in Brooklyn (e.g., Flatbush) are necessary; as well as to those that are in the process of being gentrified (Bushwick, for example) or represent authentic expatriate neighbourhoods (Coney Island). A systematic focus will have to be laid on everyday practices of its inhabitants and their discourses. Interviews, scanning the semiotic landscape as well as a systematic analysis of historical data will have to be among the data types to be collected to focus on enregisterment and practices of place-making in relation to what has recently been termed "metrolingualism" (Pennycook & Otsuji 2015) and transglossia (Sultana, Dovchin, & Pennycook 2015; García 2009, 2014; Blackledge & Creese 2010).

The fusion of theoretical concepts of place and enregisterment with a view of discourse that is constructive has illustrated that linguistic and multimodal practices in an urban space may be patterned and take on social functions to create identity and belonging. Patterned structures may be construed and realised on different levels of abstractions. Hence, urbanity is a network of values which is multidimensional and consists of a plurality of signs and agents interacting with each other. This bridging of concepts has also methodological implications for novel and fruitful ways of innovative research as the complexity of urban data demands and allows for a combination of quantitative and qualitative research methods from both sociolinguistics, linguistic landscape studies and corpus linguistics.

References

Agha, A. (2003). The social life of cultural value. *Language and Communication*, *23*(3–4), 231–273.

Agha, A. (2005). Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, *15*(1), 38–59.

The American National Corpus. (2005). 15 million words, 1990-present. Retrieved from http://www.anc.org/ *(*last accessed October 2018).

Appadurai, A. (Ed.). (1986). *The Social Life of Things. Commodities in Cultural Perspective*. Cambridge, UK: Cambridge University Press.

Augé, M. (2008 [1995]). *Non-Places: An Introduction to Supermodernity*. (2nd ed.; J. Howe, Trans.). London, UK and New York, NY: Verso.

Augé, M., & Colleyn, J. P. (2006). *The World of the Anthropologist*. Oxford, UK: Berg.

Auster, P. (2006). *The Brooklyn Follies*. New York, NY: Henry Holt and Company.

Baker, P. (2014). *Using Corpora to Analyse Gender*. London, UK: Bloomsbury.

Beal, J. C. (2009a). Enregisterment, commodification, and historical Context: "Geordie" versus "Sheffieldish." *American Speech*, *84*(2), 138–156.

Beal, J. C. (2009b). "You're not from New York City, you're from Rotherham": Dialect and identity in British Indie music. *Journal of English Linguistics*, *37*(3), 223–240.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Brookes, G., Harvey, K. & Mullany, L. (2016). 'Off to the best start?' A multimodal critique of breast and formula feeding health promotional discourse, *Gender and Language*, *10*(3), 340–363.

Blackledge, A., & Creese, A. (2010). *Multilingualism: A Critical Perspective*. London, UK and New York, NY: Continuum Press.

Blommaert, J. (2013). Semiotic and spatial scope. Towards a materialist semiotics. In M. Böck & N. Pachler (Eds.), *Multimodality and Spatial Semiosis. Communication, Meaning-making, and Learning in the Work of Gunther Kress* (pp. 29–38). New York, NY and London, UK: Routledge.

Blunt, A., & Dowling, R. (2006). *Home*. London, UK: Routledge.

Britain, D. (2017). Beyond the 'gentry aesthetic': Elites, Received Pronunciation and the dialectological gaze in England. *Social Semiotics*, *27*(3), 288–298.

Busse, B. (2010). Adverbial expressions of stance in early modern 'spoken' language. In J. Helbig (Ed.). *Anglistentag 2009 Klagenfurt – Proceedings* (pp. 47–64). Trier, Germany: WVT.

Busse, B. (2013). Genre. In P. Stockwell & S. Whiteley (Eds.) *The Cambridge Handbook of Stylistics* (pp. 103-116). Cambridge, UK: Cambridge University Press.

Busse, B. (2014). New historical stylistics. In M. Burke (Ed.) *The Routledge Handbook of Stylistics* (pp. 101–117). London, UK: Routledge.

Busse, B. (2018). Current British English: A sociolinguistic perspective. In V. Brezina, R. Love, & K. Aijmer (Eds.). *Corpus Approaches to Contemporary British Speech* (pp. 16-26). London, UK: Routledge.

Busse, B., & Warnke, I. H. (2015). Sprache im urbanen Raum. In E. Felder & A. Gardt (Eds.), *Handbuch Sprache und Wissen* (pp. 519–538). Berlin, Germany: de Gruyter.

Capote, T. (2001). *In Cold Blood*. New York, NY: Knopf Doubleday (Random House).

Carter, R. (2012). Coda: Some rubber bullet points. *Language and Literature*, *21*(1), 106–114.

Clark, H. H., & Clark, E. V. (1977). *Psychology and Language: An Introduction to Psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.

Coupland, N. (2007). *Style: Language Variation and Identity*. Cambridge, UK: Cambridge University Press.

Cresswell, T. (2004). *Place. A Short Introduction*. Oxford, UK: Blackwell.

Cresswell, T. (2006). *On the Move: Mobility in the Modern Western World*. London, UK: Routledge.

De Certeau, M. (1984). *The Practice of Everyday Life*. (S. Rendall, Trans.). Berkeley, CA: University of California Press.

Ebeling, J., & Oksefjell Ebeling, S. (2013). 'To find oneself a partner' vs. 'to find a partner': A contrastive analysis of the patterns V REFL NPindef and V NPindef in English and Norwegian. *Languages in Contrast*, *13*(2), 212–237.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of* Anthropology, *41*, 87–100.

Eckert, P., & Rickford, J. (2001). *Style and Sociolinguistic Variation*. Cambridge, UK: Cambridge University Press.

Franz, Y. (2015). *Gentrification in Neighbourhood Development. Case Studies from New York City, Berlin and Vienna*. Vienna, Austria: Vienna University Press.

Friedmann, J. (2010). Place and place-making in cities: A global perspective. *Planning Theory & Practice*, *11*(2), 149–165.

García, O. (2009). *Bilingual Education in the 21st Century: A Global Perspective*. Oxford, UK: Wiley.

García, O. (2014). Countering the dual. Transglossia, dynamic bilingualism and translanguaging in education. In R. S. Rubdy & L. Alsagoff (Eds.), *The Global-local Interface and Hybridity: Exploring Language and Identity* (pp. 100–118). Bristol, UK: Multilingual Matters.

Giles, H., & Smith, P. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. N. St. Clair (Eds.), *Language and Social Psychology* (pp. 45–65). Baltimore, MD: University Park Press.

Givón, T. (1993). *English Grammar: A Function-Based Introduction*. *Volume 1*. Amsterdam, Netherlands: John Benjamins.

Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 3–25). Amsterdam, Netherlands: John Benjamins.

Hall-Lew, L, & Stephens, N. (2012). Country Talk. *Journal of English Linguistics*, *40*(3): 256–280.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar* (3rd ed.). London, UK: Arnold.

Heidegger, M. (1993 [1977]). Building dwelling thinking. In D. F. Krell (Ed.), *Basic Writings from 'Being and Time' (1927) to 'The Task of Thinking' (1964)* (2nd rev. and expanded ed., pp. 347–363). London, UK: Routledge.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London, UK: Routledge.

Jacobs, J. (1962). *The Death and Life of Great American Cities*. New York, NY: Random House.

Jaworski, A., & Thurlow, C. (2010). Introducing semiotic landscapes. In A. Jaworksi & C. Thurlow (Eds.). *Semiotic Landscapes: Language, Image, Space* (pp. 1–40). London, UK: Continuum.

Johnstone, B. (2009). Pittsburghese shirts: Commodification and the enregisterment of an urban dialect. *American Speech*, *84*(2), 157–175.

Johnstone, B. (2010). Language and geographical space. In P. Auer & J.E. Schmidt (Eds.). *Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods* (pp. 1–18). Berlin, Germany: de Gruyter.

Johnstone, B. (2011). Dialect enregisterment in performance. *Journal of Sociolinguistics*, *15*(5), 657–679.

Johnstone, B. (2014). '100% Authentic Pittsburgh': Sociolinguistic authenticity and the linguistics of particularity. In V. Lacoste, J. Leimgruber, & T. Breyer (Eds.). *Indexing Authenticity: Sociolinguistic Perspectives* (pp. 97-112). Berlin, Germany: De Gruyter.

Johnstone, B, Andrus, J., & Danielson, A. E. (2006). Mobility, indexicality, and the enregisterment of “Pittsburghese.” *Journal of English Linguistics*, *34*(2), 77–104.

Jordan, M. P. (1998). The power of negation in English. Text, context and relevance. *Journal of Pragmatics*, *29*(6), 705–752.

Joyce, P. (1991). *Visions of the People: Industrial England and the Question of Class 1848–1914*. Cambridge, UK: Cambridge University Press.

Kallen, J. (2010). Changing landscapes: Language, space and policy in the Dublin linguistic landscape. In A. Jaworksi & C. Thurlow (Eds.). *Semiotic Landscapes: Language, Image, Space* (pp. 41–58). London, UK: Continuum.

Lacoste, V., Leimgruber, J., & Breyer, T. (Eds.). (2014). *Indexing Authenticity: Sociolinguistic Perspectives*. Berlin, Germany: De Gruyter.

Lakoff, G. (2004). *Don’t Think of an Elephant! Know your Values and Frame the Debate: The Essential Guide for Progressives*. White River Junction, VT: Chelsea Green Publishing.

Lakoff, G. (2006, 14 February). Simple framing: An introduction to framing and its uses in politics. Retrieved from https://georgelakoff.com/writings/rockridge-institute/ (last accessed August 2018).

Leech, G. (1983). *Principles of Pragmatics*. London and New York, NY: Longman.

Lefebvre, H. (1974). *The Production of Space.* (D. Nichsolson Smith, Trans.). Malden, MA: Blackwell.

Le Page, R. B., & Tabouret-Keller, A. (1985). *Acts of Identity: Creole-based Approaches to Language and Ethnicity*. Cambridge, UK: Cambridge University Press Archive.

Li, W. (2011). Moment analysis and translanguaging space: Discursive construction of identities by multilingual Chinese youth in Britain. *Journal of Pragmatics*, *43*(5), 1222–1235.

Mahlberg, M. (2013). *Corpus Stylistics and Dickens’s Fiction*. New York, NY: Routledge.

Maher, J. C. (2005). Metroethnicity, language and the principle of Cool. *International Journal of the Sociology of Language*, *175/176*, 83–102.

Mair, C. (2013). The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide*, *34*(3), 253–278.

Massey, D. (2005). *For Space*. London, UK: Sage.

McIlvenny, P., Broth, M., & Haddington, P. (2009). Communicating place, space and mobility. *Journal of Pragmatics*, *41*(10), 1879–1886.

Moore, E. (2012). The social life of style. *Language and Literature*, *21*(1), 66–83.

OED Online. (2018). Oxford University Press. http://www.oed.com (last accessed October 16, 2018).

Osman, S. (2012). *The Invention of Brownstone Brooklyn: Gentrification and the Search for Authenticity in Postwar New York*. Oxford, UK: Oxford University Press.

Otsuji, E., & Pennycook, A. (2010). Metrolingualism: Fixity, fluidity and language in flux. *International Journal of Multilingualism*, 7(3), 240-254.

Partington, A., Duguid, A. & Taylor, C. (2013). *Patterns and Meanings in Discourse Theory and Practice in Corpus-assisted Discourse Studies (CADS)*. Amsterdam, Netherlands: John Benjamins.

Pennycook, A., & Otsuji, E. (2015). *Metrolingualism: Language in the City*. London, UK: Routledge.

Popik, B. (2006, April 18). “How sweet it is!” & Brooklyn street signs [Blog post]. Retrieved from http://www.barrypopik.com/index.php/new_york_city/entry/how_sweet_it_is_brooklyn_street_signs/ (last accessed October 2018).

Relph, E. (1976). *Place and Placelessness.* London, UK: Pion.

Sassen, S. (1994). *Cities in a World Economy*. Thousand Oaks, CA: Pine Forge Press.

Searle, J. R. (1976). *A Taxonomy of Illocutionary Acts*. Trier, Germany: Linguistic Agency, University of Trier (LAUT).

Searle, J. R. (1995). *The Construction of Social Reality*. New York, NY: Free Press.

Silverstein, M. (1993). Metapragmatic discourse and metapragmatic function. In J. Lucy (Ed.), *Reflexive Language: Reported Speech and Metapragmatics* (pp. 33–58). Cambridge, UK: Cambridge University Press.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, *23*(3–4), 193–229.

Simpson, P. (1999). Language, culture and identity: With (another) look at accents in pop and rock singing. *Multilingua*, *18*(4), 343–367.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, *7*(2), 215–44.

Stubbs, M. (2013). Sequence and order: The neo-Firthian tradition of corpus semantics. In H. Hasselgård, J. Ebeling, & S. Oksefjell Ebeling (Eds.) *Corpus Perspectives on Patterns of Lexis* (pp. 13–34). Amsterdam: John Benjamins.

Sultana, S., Dovchin, S. & Pennycook, A. (2015). Transglossic language practices of young adults in Bangladesh and Mongolia. *International Journal of Multilingualism*, *12*(1), 93—108.

Thompson, G. (1996). *Introducing Functional Grammar*. London, UK: Arnold.

United States Census Bureau. (2007). *County and City Data Book: 2007. A Statistical Abstract Supplement*. Washington, DC: US Census Bureau.

Venturi, R. (1966). *Complexity and Contradiction in Architecture.* New York: The Museum of Modern Art Press.

Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies*, *30*(6), 1024—1054.

Warnke, I. (2013). Diskurs als Praxis und Arrangement – Zum Status von Konstruktion und Repräsentation in der Diskurslinguistik. In W. Viehöver, R. Keller & W. Schneider (Eds.), *Diskurs – Sprache – Wissen. Interdisziplinäre Beiträge zum Verhältnis von Sprache und Wissen in der Diskursforschung* (pp. 97–117). Wiesbaden, Germany: Springer.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge, UK: Cambridge University Press.

Zhang, Q. (2005). A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity. *Language in Society*, *34*(3), 431—466.

Zukin, S. (2010). *Naked City: The Death and Life of Authentic Urban Places*. Oxford, UK: Oxford University Press.

Dagmar Deuber and Eva Canan Hänsel

The English of current Caribbean newspapers

American, British, in between or neither?

Abstract: The present paper analyses newspaper corpora from Jamaica and three small island nations in the Caribbean, namely St. Kitts & Nevis, Dominica and St. Vincent & the Grenadines, with the aim of shedding more light on the issue of American influence in written Standard English in the Caribbean. Spelling, lexis and selected grammatical features are analysed. This is complemented by investigating the perspectives of newspaper staff and readers. The results for spelling are mixed and can be interpreted to show that language use in the small Caribbean countries tends to be more susceptible to American influence. Concerning lexis, the findings are limited but suggest that several lexical items associated with American English are well established across the different Caribbean countries. In terms of grammatical forms and constructions, all the Caribbean corpora are characterised by a distinctive tendency towards formal variants. Finally, the study suggests that next to exonormative influences, endonormative attitudes need to be considered when assessing the current state of written Standard English in the Caribbean.[1]

1 Introduction

The Anglophone Caribbean has been defined as comprising those of the Caribbean islands that have English as an official language as well as the mainland territories of Belize and Guyana; altogether there are twelve independent nations, all formerly part of the British Empire, and six dependencies, five of the UK and one of the USA (see Allsopp 1996: xvii–xix; Deuber 2014: 4; Winford 1991: 565–

[1] The revised version of this paper has benefited from comments by the reviewers as well as Viola Wiegand; any remaining shortcomings are our own responsibility. While producing this version we received financial support for our research from the German Research Foundation (DFG; grant number: DE 2324/1-1).

Dagmar Deuber, University of Münster, deuber@uni-muenster.de
Eva Canan Hänsel, University of Münster, evahaensel@uni-muenster.de

https://doi.org/10.1515/9783110489071-003

568). Typical of this region is a range of spoken language use comprising acrolectal (English), mesolectal (intermediate Creole) and basilectal (conservative Creole) varieties, though some territories lack the last-named type of variety (see Winford 1993: 4 for an overview). Languages other than English and English-based Creole are now commonly spoken only in Dominica and St. Lucia (French-based Creole) and Belize (Spanish). The spoken acrolect comprises a formal standard (Allsopp 1996: lvi) and more informal educated usage (Deuber 2014). Written language use remains dominated by Standard English, although in recent times there has been a surge of writing in Creole in informal genres of computer-mediated communication (see e.g. Hinrichs 2006; Mair 2011; Oenbring 2013).

Apart from the Caribbean-wide documentation of lexis in Allsopp (1996), linguistic work on acrolectal varieties in the Caribbean has mainly focused on Jamaica and Trinidad & Tobago,[2] the two largest countries in terms of population with 3 and 1.2 million inhabitants, respectively (population figures according to Central Intelligence Agency 2018); the few studies on acrolectal varieties elsewhere include a first approach to Bahamian Standard English (Bruckmaier & Hackert 2011). This situation is to a great extent related to the development of corpora for English in the Caribbean in the framework of the International Corpus of English (ICE), as many of the studies in question have been based on data from these corpora. ICE currently includes three Caribbean countries, namely Jamaica, Trinidad & Tobago and the Bahamas. ICE-Jamaica, for a long time the only ICE project in the region, was completed in 2009. ICE-Trinidad & Tobago and ICE-Bahamas (see Deuber 2010b and Hackert 2010, respectively) are more recent and still ongoing projects.

Research so far has devoted considerable attention to spoken English (e.g. for Jamaica: Deuber 2009a; Gut 2011; Irvine 1994, 2004, 2008; Jantos 2010; Mair 2009; Rosenfelder 2009; Sand 1999; for Trinidad: Deuber 2009b, 2010a; for both: Deuber 2014). Written English in the Caribbean, not being as distinctive from varieties of English elsewhere, has received less attention in recent studies (apart from the case of computer-mediated communication, as mentioned above). There has, however, been a long-standing interest in student writing (see e.g. Christie 1989; Craig 1997; Mair 2002). Moreover, newspaper English was one of the subjects in Sand's (1999) study on Jamaica and has more recently been analysed by Bruckmaier and Hackert (2011) for the Bahamas and by Hänsel and Deuber (2013) for Trinidad & Tobago. These three studies all find American influences in the

[2] Much of the research on the latter has specifically focused on Trinidad, by far the larger of the two islands.

area of lexis, while spelling is reported to be predominantly British-oriented. Bruckmaier and Hackert (2011) additionally identify several areas of grammar where Bahamian newspaper English follows American norms.

As Milroy and Milroy (1999) emphasise, the notion of Standard English has an important ideological component. Generally a standard language ideology can be defined as "the widely circulating belief system that promotes the standard language as better than other varieties and the only correct option" (Curzan 2014: 119). Sand (1999: 180) in her pioneering study of Jamaican English in fact drew attention to the importance of such attitudes in determining what counts as Standard English. However, to date only few studies of Caribbean standard varieties have considered this aspect. These include Deuber and Leung (2013) on newscasters' accents in Trinidad. Sand herself has more recently presented results of an attitude study based on radio and newspaper data from ICE-Jamaica. Certain parts of her findings suggest that "[f]or many Jamaicans, 'good English' or 'proper English' in writing are still associated with the former exonormative standard which is believed to be British English" (Sand 2011: 169). In a language attitude study in Jamaica as well as Trinidad, Deuber (2013) found both groups of informants to be divided between accepting the notion of an endonormative standard and upholding the notion of an exonormative one, with the latter often specified as British, though some informants pointed out a mixture of British and American influences.

Findings so far suggest that there may be a discrepancy between use and attitudes when it comes to the influence of American English on Standard English in the Caribbean. To study this issue more closely consideration of different aspects is therefore necessary. This is what the present chapter sets out to do. It links up with the tradition of research on English in the Caribbean as outlined above in that it deals with newspaper English and in that it includes Jamaica as one of the countries to be studied, but it also breaks new ground by considering three small island nations: St. Kitts & Nevis (pop. 53,000), Dominica (pop. 74,000), and St. Vincent & the Grenadines (pop. 102,000) (population figures according to Central Intelligence Agency 2018). We analyse the frequency of spellings, lexical items and selected grammatical constructions relevant to the issue at hand in newspaper corpora for the four countries. To complement this we investigate the perspectives of newspaper staff and readers. Telephone interviews with editors or other staff members were conducted to find out about regulations or guidelines concerning varieties of English at the different newspapers and comments posted online were analysed with regard to readers' attitudes towards spellings, lexical items and grammatical constructions associated with American

English (or considered by the author of the comment to be so). The following section (2) will provide details of the data and methods used. Section 3 will present the results of the corpus analyses, while Section 4 will focus on the attitudes of newspaper staff and readers. The results of both these sections will be drawn together in the final discussion in Section 5. The last section (6) will give our conclusions.

2 Data and methods

The present study uses newspaper corpora for Jamaica (JAM), St. Kitts & Nevis (SKN), Dominica (DMA) and St. Vincent & the Grenadines (SVG) that are of equal size and composition: about 180,000 words in total, of which about 120,000 in the category of reportage and about 60,000 in the category of editorials. While editorials also address international issues that are of interest to the Caribbean region (e.g., the Olympic Games, implications of the outcome of the 2012 US elections for Jamaica), the articles in the reportage section almost exclusively address local, i.e. national issues (e.g., local politics, sports, culture). Only very rarely do the reportage articles also address regional or international issues if they are of relevance to the Caribbean region (e.g., the Falkland Islands referendum on whether to remain a British Overseas Territory). No articles that were labelled as having been provided by international news agencies were included in the data. The corpora were mainly compiled from web editions of established print newspapers. However, in the case of Dominica only one print newspaper offered a web edition and data from news websites were therefore added. The total number of sources is four each for St. Kitts & Nevis and for Dominica, three for Jamaica, and two for St. Vincent & the Grenadines.[3] The articles included in the corpora are mostly from 2012 with some from 2011 or 2013. The corpora were part-of-speech tagged with the C7 tagset using CLAWS (Rayson & Garside 1998).[4] For comparison, the press sections of the British English 2006 (BE06) and American English 2006 (AE06) corpora were used; they were accessed at https://cqpweb.

[3] In Appendix A, we specify the names and URLs of the newspapers and news websites from which the corpus data were drawn. However, in order to preserve a degree of confidentiality with regard to the exact provenance of the information provided by the newspaper staff we interviewed (see Section 4.1), we refer to the individual sources by Roman numerals rather than by the newspapers' names.

[4] No manual post-editing was done.

lancs.ac.uk/ (Hardie 2012).[5] These corpora, which are also annotated with C7 part-of-speech tags, replicate the Brown family design with data from around 2006 (see Baker 2009 on BE06). The press sections consist of 88 texts of about 2,000 words each, thus comprising about 176,000 words in total, with the texts distributed across 3 categories: 44 reports, 27 editorials and 17 reviews. While not matching the Caribbean corpora exactly, they were considered reasonably comparable for present purposes.

The corpora were subjected to quantitative analyses using the WordSmith 5.0 Tools concordancer (Scott 2008) for the Caribbean corpora and the CQPweb query tool for BE06 and AE06, respectively, with irrelevant or mistagged items sorted out manually. The analyses cover spellings and lexical items associated with American English and selected grammatical features. The grammatical features studied are *that* versus *which* in relative clauses, the *be*-passive as well as verb and negative contractions; all these are phenomena for which Leech et al. (2009) have documented substantial differences between American and British English in frequency of use.

The interviews were conducted among editors or other staff members of as many of the newspapers and news websites used for the corpora as could be reached (all but one of the Jamaican and one of the Dominican sources). The interviewer (the second author of this paper) followed a predetermined schedule with a specific set of questions and follow-up questions, which can be found in Appendix B. The interviews were conducted via telephone as field trips to the four countries under study were not feasible.[6] It has been suggested that this mode of administration may negatively affect the data, but this seems to apply mainly when sensitive issues are being addressed and/or interviews are of long duration (Bryman 2012: 215). None of this was the case in the present study (the interviews lasted no more than 15 minutes), so conducting the interviews by telephone could be expected to be unproblematic. However, it needs to be acknowledged that while most interviewees were happy to answer the questions, one of them was rather sceptical.

The readers' perspective was investigated through comments posted online, i.e. by means of a societal treatment type of approach to language attitudes (see Garrett 2010). An advantage of this approach is that language attitudes were

[5] However, the results from these corpora were normalised not on the basis of the word counts provided by CQPweb, as these are higher than traditional word counts (Andrew Hardie, p.c.), but rather on the basis of word counts according to Microsoft Word kindly provided by Paul Baker (p.c.) and Amanda Potts (p.c.) for BE06 and AE06, respectively.

[6] Two editors preferred to answer the questions via e-mail.

investigated within the very same context as language use, as comments were taken from a website from which corpus data was also drawn. Furthermore, these attitudes were articulated spontaneously and not elicited by a researcher. However, a limitation is that there is a bias towards a certain type of respondent, namely those who are inclined to post comments in the first place, and among them those who care sufficiently about language matters to address them. As not all news websites have a comment section, this part of the study analyses, by way of example, the comments found on one news website, Dominica News Online, where quite a lot of language-related discussion was found. The Google search engine was used to search for the terms *American* and *English* on the website (search string: site:http://dominicanewsonline.com american english).[7] All language-related items were extracted from the total results and then classified based on the type of material that triggered the comment as well as on the aspect of language use addressed.

3 Analyses of the corpora

This section deals first with British versus American spellings (3.1), followed by lexical items (3.2). We then proceed to the grammatical features (3.3); the results for these will be considered in the light of research on grammatical variation and change in present-day British and American English based on the Brown family of corpora.

3.1 Spelling

The analysis of spelling takes into account most of the major systematic differences between American and British English as described by Tottie (2002: 9–11).[8] For each of these systematic differences all relevant words attested in the corpora were searched for (see Appendix C for examples). The analysis also has a "miscellaneous" category comprising a couple of word pairs that exhibit idiosyncratic differences (see also Appendix C). Excluded from the analysis are all proper names.

[7] The search was carried out on 23 April 2014.

[8] Not included are <log> versus <logue> in words such as *dialog(ue)* (there are only 4 attestations of the former in the AE06 press data but 12 of the latter) and <e> versus <ae> (or <oe>) in words such as *(a)esthetic* (these hardly occur in the data).

The results for each category are presented in Appendix C. Note that <-iz-> versus <-is-> was analysed because the British and American data contrast with respect to this feature in the same way as they do for the other features. However, as the Caribbean corpora mostly show much higher frequencies of <-iz-> relative to <-is-> than of the other American spellings, and as <-iz-> is a possible spelling in British English as well, this category has not been included in the overall results.

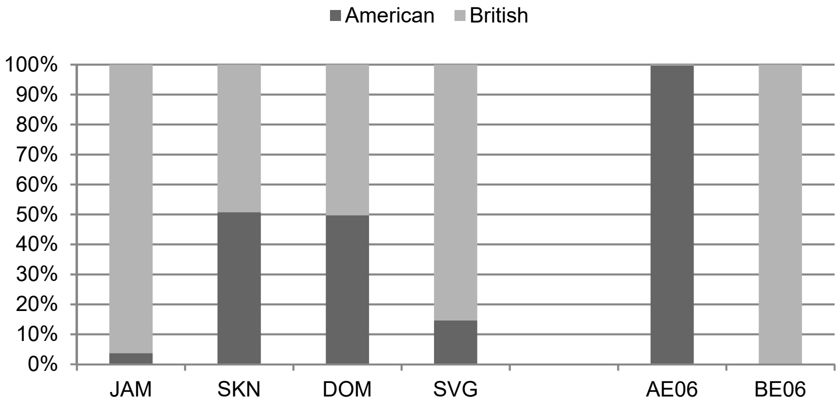


Fig. 1: Overall results for American and British spelling variants excluding <-iz-> versus <-is->.

The overall results for each corpus are shown in Figure 1. In Figure 2 they are broken down by individual data sources. As can be seen, Jamaican newspaper English is very closely aligned with British English. British spellings are also in a clear majority in the data from St. Vincent & the Grenadines but the proportion of American spellings is appreciably higher than in the Jamaican data (15% versus 4%). In the data from St. Kitts & Nevis and Dominica spelling practices are very mixed, with proportions of American spellings of 51% and 50%, respectively. In the case of Dominica the type of source is an apparent factor: sources I–III, which show high proportions of American spellings, are all news websites, whereas source IV, where a low proportion of American spellings was found, is an online edition of a print newspaper. However, there is also a wide range of variation among the four sources from St. Kitts & Nevis although they are all online editions of print newspapers.

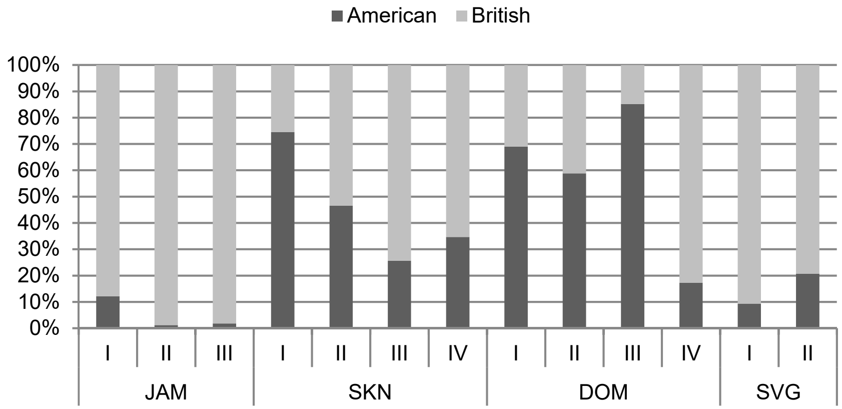


Fig. 2: Overall results for American and British spelling variants excluding <-iz-> versus <-is-> by individual data sources.

That the Jamaican corpus is overall closest to British English in terms of spelling could be taken to mean that at the present stage Jamaica has a stronger orientation towards British English than the three small countries under investigation. However, it should also be considered in this connection that evidence for a British orientation of Jamaican English "seems to be disappearing outside the relatively firmly regulated area of spelling" (Mair 2009: 59), and that Jamaica's education system explicitly recognises a 'Standard Jamaican English' (see Deuber 2013: 121–122, 2014: 39). Therefore, the use of spellings historically inherited from British English may also be due to an orientation to what has become established practice in Jamaica rather than British English as such.

3.2 Lexical items

While there is a considerable number of lexical contrasts between British and American English (see e.g. Kövecses 2000: Chapter 10; Tottie 2002: Chapter 5), the number of lexical items for which single-genre corpora of a rather small size such as the present ones can supply sufficient attestations is obviously limited. On the other hand, though, those items that are attested with sufficient frequency in these corpora belong to the most recurrent and relevant ones in the English of Caribbean newspapers of those that show a contrast between British and American English. Whether such a contrast exists is also something that had to be confirmed in the analysis: as there has been a continuous tendency for lexical items to spread from American to British English (Kövecses 2000: 248–252; Peters 2001; Trudgill 1998),[9] only items were chosen which are also sufficiently attested in BE06 and AE06 and for which these corpora display a fairly clear-cut distinction. A total of 9 word pairs was thus identified as suitable for analysis. All proper names are again excluded.

The results are shown in Table 1. Barring some inconclusive results due to low numbers, in seven out of the nine cases analysed the Caribbean corpora tend to show a common preference for the item associated with American English (*math, cell(ullar) (tele)phone, transportation*, *garbage, vacation, student, principal*), while in the remaining two cases the item associated with British English (*trade(s) union(ist), towards*) is the one mainly used in the Caribbean corpora. Thus, beyond the evident American influence a shared pattern uniting the different Caribbean corpora emerges.

Tab. 1: American and British lexical items (raw numbers).

	JAM	**SKN**	**DMA**	**SVG**	**AE06**	**BE06**
math	17	6	2	6	11	0
maths	4	0	0	3	0	10
cell phone(s)/cellphone(s)/cellular (tele)phones	7	3	5	5	5	1
mobile (tele)phone(s)	0	0	1	0	1	21
transportation	13	1	11	7	18	0
transport[a]	3	0	0	2	1	25
garbage	18	14	4	15	3	1
rubbish	4	1	3	1	1	9
vacation(s)	2	5	3	13	4	0
holiday(s)[b]	2	2	2	3	0	8
student(s)[c]	157	159	133	164	58	10
pupil(s)[d]	1	1	0	5	0	31
principal(s)[e]	53	12	19	30	13	0
headmaster(s)/head teacher(s)	1	0	1	5	1	5
labor union(s)/-ist(s)	0	0	0	0	5	0
labour union(s)-ist(s)	0	1	1	0	0	0
trade(s) union(s)/-ist(s)	7	3	13	5	1	3

[9] The reverse process is also attested but much rarer (Kövecses 2000: 150–151).

	JAM	SKN	DMA	SVG	AE06	BE06
toward	3	2	7	2	54	1
towards	31	41	48	53	1	25

[a] Including only nouns.

[b] In the sense of ‘vacation’ only.

[c] Excluding instances referring to students at tertiary level.

[d] Excluding instances referring to the part of the eye.

[e] In the sense of ‘headmaster’ only.

3.3 Grammatical features

This section presents and discusses our analyses of *that* versus *which* in relative clauses, the *be*-passive and contractions, in the order mentioned.

3.3.1 *That* versus *which* in relative clauses

Concerning the use of relativisers and specifically the choice between *that* and *which* Leech et al. (2009: 228–230), analysing the Brown family of corpora, found a remarkable trend of divergence between American English and British English in the thirty-year period covered by these corpora: the relativiser *that* was already more frequent in American than in British English in the 1960s and its frequency increased steeply in American English between the 1960s and 1990s, but only moderately in British English, so that the gap between the two widened. These findings have been confirmed in the more detailed analysis by Hinrichs, Szmrecsanyi and Bohmann (2015). They conclude that a trend towards *that* as the more colloquial variant has been reinforced by the prescriptive recommendation that *which* should not be used in restrictive relative clauses and they therefore describe this change as a case of “institutionally backed colloquialization-*cum*-Americanization” (Hinrichs, Szmrecsanyi, & Bohmann 2015: 806).

The present analysis is concerned with *that* and *which* in those contexts where they are in competition; (1) and (2) below are examples.

(1) Why didn’t he attack and vilify the electoral system and make the **same allegations which** he is making now? (SKN)

(2) She urged the parents to step in and assist **the preparatory schools that** are struggling. (JAM)

Traditionally a distinction is made between restrictive relative clauses, which allow both, and non-restrictive relative clauses, where normally *which* is used, though *that* may occasionally occur (Huddleston & Pullum 2002: 1052). However, this distinction is not unproblematic. Huddleston and Pullum (2002) rather refer to the two types as 'integrated' and 'supplementary' relative clauses because the contrast is often not a matter of whether the relative clause restricts the reference of the antecedent but of whether its content is presented as an integral part of the message, so that the same relative clause may in fact have integrated as well as supplementary readings (Huddleston & Pullum 2002: 1064–1065). In written Standard English supplementary relative clauses are usually separated from the rest of the sentence by commas or occasionally other punctuation marks. This is not always the case though (Huddleston & Pullum 2002: 1056). In their absence, however, it can be difficult to unambiguously determine that a relative clause is supplementary rather than integrated unless the antecedent is an adjective phrase, verb phrase, or clause and not a noun phrase (Huddleston & Pullum 2002: 1052). In view of this, the present analysis excludes all relative clauses marked off by commas or other punctuation marks (see example 3), and beyond that any relative clause whose antecedent is an adjective phrase, verb phrase, or clause (4), but otherwise no attempt was made to identify supplementary relative clauses not marked off by punctuation marks. A few further contexts were excluded because they allow only either *that* or *which* (Huddleston & Pullum 2002: 1052–1054): those where the relativised element is complement of a preposition and the preposition is fronted, i.e. the pied piping construction (*which* but not *that*) (5), those where the antecedent is the pronoun *that* (*which* but not *that*) (6), those where the antecedent is human (*that* but not *which*) (7), and those where the relativised element functions as adjunct (*that* but not *which*) (8). All instances to be excluded from the counts were identified manually.

(3) The most recent **"bomb threat", which** was called in at the Ministry of Agriculture on Tuesday, has once again raised a number of questions [...]. (SVG)

(4) Gonsalves also explained that there was a social problem in that **those who opted to make a living in the hills often spent long periods of time in the hills which** had the potential to affect their family life. (SVG)

(5) It is **a matter to which** the authorities must give serious consideration. (SVG)

(6) Are we too afraid to accept **that which** is true, that which is real in live and in living colour in the society in which we have to live and die? (SVG)

(7) He was **a man that** you could take any pattern from. (JAM)

(8) Prime Minister Skerrit [...] pointed out that it [the constitution] makes provision for a letter of resignation from the President to take effect on **the day that** the letter reaches the speaker of the House. (DMA)

The results of the present analysis are displayed in Figure 3. According to the results for AE06 and BE06, the gap between American and British English has narrowed again since the 1990s. Whereas a comparable analysis by Hinrichs, Szmrecsanyi and Bohmann (2015: 828) yielded proportions of *that* of about 44% and 97% in the press sections of F-LOB (British English, 1990s) and Frown (American English, 1990s), respectively,[10] the proportions in BE06 and AE06 are 75% and 99%, respectively. However, considering that *which* remains a viable option in integrated relative clauses in British press language, in contrast to its near absence in the American data, relativiser use is still potentially a useful indicator of whether current Caribbean newspaper English is aligned rather with American English or with British English. It turns out that the Caribbean corpora are clearly not aligned with American English as the relativiser *which* is used to a rather high degree. The frequency of *which* is even significantly higher in all four Caribbean corpora than in BE06.[11] Proportions of the use of *which* range from about 33% in the case of Jamaica to about 42% in the case of Dominica, as compared to 25% in BE06. Therefore, the Caribbean corpora can hardly be said to be aligned with current British English either.

[10] Their exclusions are essentially the same as in the present study. The only exceptions are relative clauses not marked by punctuation marks whose antecedent is an adjective phrase, verb phrase or clause and relative clauses whose antecedent is the pronoun *that*, which were excluded only in the present study, but the number of these is small. Note in addition that relative *that* and *which* were identified automatically rather than manually in their study as they used versions of the corpora annotated with the C8 tagset, which, in contrast to the C7 tagset, has separate tags for *that* and *which* in relativiser function.

[11] All differences between each of the Caribbean corpora and BE06 are significant according to pairwise chi square tests (JAM vs. BE06: χ^2 (1, n = 1128) = 7.74, p = .005; SKN vs. BE06: χ^2 (1, n = 1118) = 8.93, p = .003; DMA vs. BE06: χ^2 (1, n = 1163) = 30.73, p = < .001; SVG vs. BE06: χ^2 (1, n = 977) = 19.89, p < .001).

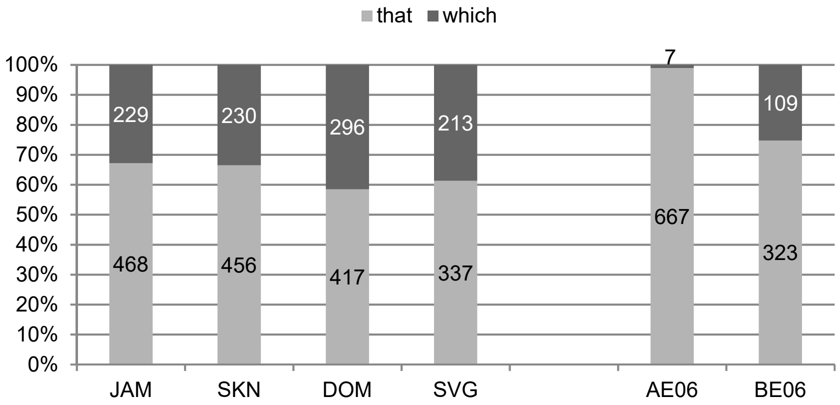


Fig. 3: Relative *that* versus *which* (box sizes indicate proportions, labels indicate raw numbers).

3.3.2 The *be*-passive

The *be*-passive has seen a decline in American and British English which, like the rise of relative *that,* has been attributed to both colloquialisation and prescriptive pressure (Mair 2006: 190–191; Leech et al. 2009: 148–152; Smith & Leech 2013: 92–95). American English has also been in the lead of this change.

Smith and Leech (2013: 94) report for British English that non-finite passives have been relatively resistant to the trend of decline compared to finite ones. They note that examples cited e.g. in usage guides are mostly finite and therefore suspect that non-finite passives are less salient as passive forms (Smith & Leech 2013: 95). The present analysis is restricted to passives in finite verb phrases. The corpora were searched for forms of *be* (*am*, *are* and *is*, including their contracted variants, *was, were, be, been*) followed by a past participle (i.e. an item tagged as VVN, VDN or VHN) with up to four words in between. A typical example of a passive form retrieved in this way is given in (9).

(9) These territorial winners **were** then **judged** by a panel of prominent Caribbean persons including former Carib Chief of Dominica Charles Williams. (DMA)

Non-finite passives as in (10) were then excluded manually. Also excluded were, of course, irrelevant instances, where a form of *be* and a past participle happened to occur within the specified distance without belonging together, as in (11).

(10) According to him, paying VAT at the port has been a crippling factor to his operations since thousands of dollars have to **be paid** upfront for tax which is not recovered.

(11) They **are** feeling the squeeze **brought** about by the global economic crisis […]. (SVG)

Due to the high degree of overlap and ambiguity that exists between adjectival and verbal uses of past participles we generally relied on the past participle tags without further manual identification of adjectival uses, which was also the procedure in the previous studies cited above. Only instances where the past participle is *gone* or where it is *drunk* and the subject is human were excluded as these cannot be passive.

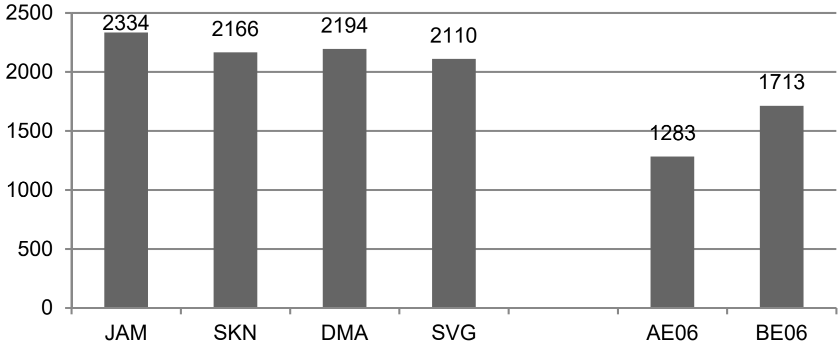


Fig. 4: The *be*-passive (passives in finite verb phrases only; frequencies normalised to 180,000 words).

The results, shown in Figure 4, reveal a major contrast between American and British English with respect to the frequency of finite *be*-passives in press language. The Caribbean corpora, meanwhile, are again in a category of their own, with considerably higher frequencies than even in British English.

3.3.3 Contractions

Contractions are a paradigm case of the colloquialisation trend observed in standard written English in research based on the Brown family of corpora. Like relativisers and the *be*-passive they have been found to show not only diachronic change, in this case a substantial increase, but also regional variation, with American English exhibiting a greater tendency towards their use.

The corpora were searched for contracted forms of *am, are, is, have, has, had, will* and *would* as well as for negative contractions involving *n't*,[12] as in the following two examples:

(12) **He's** to return to court on February 19 on a charge of threatening a witness. (JAM)
(13) Violence against Women is a topic that **isn't** always given the necessary attention [...]. (SVG)

The present results as shown in Figure 5 provide evidence of a continuing contrast between American and British English but above all they reveal a major difference between the data for these two varieties and the Caribbean data, where far lower frequencies of contractions are consistently seen.

[12] In the searches for contracted forms of *is* and *has* the corresponding tags, VBZ and VHZ, respectively, were used. Instances of the genitive marker -*s* mistagged as a verb form were excluded manually. No mistagged verb forms were found among the items tagged as the genitive marker -*s*.

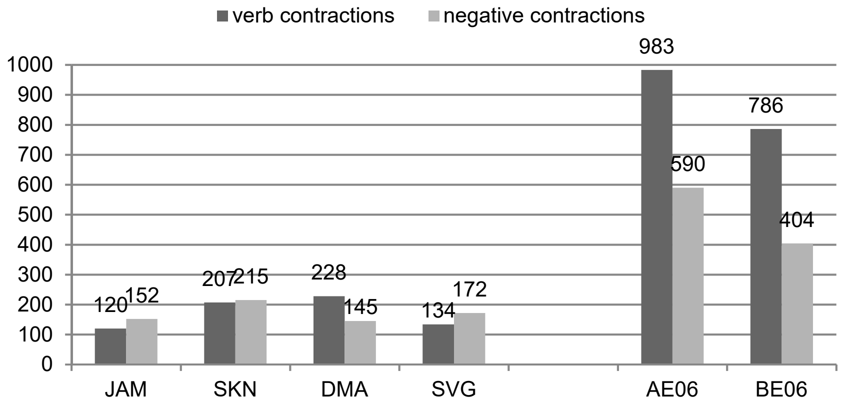


Fig. 5: Contractions (frequencies normalised to 180,000 words).

A “pronounced formality” of Jamaican newspaper language was already observed by Sand (1999: 109). The present findings for grammatical features have provided evidence that a less colloquial writing style than in British and especially American English not only continues to be characteristic of Jamaican newspaper language today but is equally a feature of the newspapers of the small Caribbean island nations included in this study.

4 Newspaper staff and readers’ perspectives

This section focuses on the perspectives of newspaper staff and readers. It first reports the findings of the interviews (4.1) and then goes on to present the analysis of the comments posted on Dominica News Online (4.2). In presenting these perspectives we will also consider how they relate to the findings from the corpora presented above (Section 3).

4.1 Newspaper staff’s perspectives

The telephone interviews included questions about the background of the writers contributing to each newspaper or news website and about regulations or preferences concerning varieties of English. The main focus was on regulations concerning British and American spellings and lexical items. Additionally, the newspaper staff were asked whether they had style guides that specify which variety of English should be used. Further questions concerned the use of spell checkers and dictionaries as well as news agency articles. Details of the questions and responses are provided in Appendix B. The labelling of the sources by Roman numbers in the Appendix as well as the following discussion of the results corresponds to Figure 2.

With regard to the writers the different sources have a lot in common. Most of the writers are stated to be from the country in question. Those who are not are in several cases stated to be from elsewhere in the Caribbean. The answer that some have studied and/or lived abroad was quite common but only in the case of some of the sources from small countries was it also stated that contributors currently lived abroad (sources II and III from Dominica and source II from St. Vincent & the Grenadines). If a specific country abroad was mentioned it was most often the USA.

The question of whether regulations concerning varieties of English were specified in a style guide was affirmed by the editors of source II from Jamaica and source II from St. Vincent & the Grenadines. The style guides at their newspapers were based on British rules. The owner of source II from Dominica stated that he is considering designing a style guide but that he is still struggling with whether to stick to the British model or whether to also allow American spellings. However, he was certain that the style guide would not prescribe American English only. All other respondents reported that they did not have a style guide.

Concerning spelling two groups can be distinguished among the answers. The first, larger group consists of those sources where it was stated in the interviews that British spelling is supposed to be used. These are the sources from Jamaica and St. Vincent & the Grenadines, sources III and IV from St. Kitts & Nevis as well as source I from Dominica. The second, smaller group includes three sources (sources I and II from St. Kitts & Nevis and source III from Dominica) that allow both British and American spelling; to these can be added a fourth (source II from Dominica) that is considering doing so due to an observed increasing influence of American English on English in Dominica.

In one case the interview statement on spelling and the result from the analysis of the corpus texts for the source in question are clearly at odds: according to the interview, source I from Dominica uses British spelling but the proportion of American spellings in the corpus data is as high as 69%. Only a short interview was obtained with this source and the information given may not be as reliable as in the other cases. Apart from this case there is a fair degree of correspondence between stated editorial practice and the spelling practices evident in the corpus data: in the first of the two groups distinguished above the proportion of American spellings ranges from 1% (Jamaican source II) to 35% (source IV from St. Kitts & Nevis) and in the second from 47% (source II from St. Kitts & Nevis) to 85% (Dominican source III).

Whereas the corpus results showed that lexical items associated with American English are more widely used than American spellings, the answers to the question concerning vocabulary were in many cases the same as for spelling. Only somewhat more acceptance of American English was expressed: in the cases of one source from Jamaica and one from St. Vincent & the Grenadines the answer is in favour of British English but less categorically so than in the case of spelling, and the Dominican source that struggles with the question of whether to explicitly allow American spelling leaves writers free to choose between British and American vocabulary. The corpus data for lexical items being less extensive, the degree of correspondence between stated and actual practice cannot be assessed quite as well as for spelling. However, it is noteworthy that although all corpora showed a preference for several common lexical items associated with American English, the sources were divided in terms of their statements whether staff should use British or American English. Accordingly, it seems safe to say that there is a certain gap between stated and actual practice.

4.2 Readers' perspectives

The language-related items identified through the Google search on Dominica News Online (DNO) mentioned in Section 2 were all from readers' comments. In terms of the trigger for the comment, three main types were identified: comments on an aspect of language use in an article on DNO, comments on an aspect of language use in another reader's comment, and comments on the language use of somebody outside the context of DNO. Comments of the last type were mostly about American accents and have not been considered for present purposes. In a few further cases an aspect of American English was brought up in the course of a discussion that had been triggered by another related matter without it having been used in any specific instance being discussed. The discussion below will proceed from spelling to lexical items and finally to grammatical features as addressed in the comments while also taking the different types of triggers into account.

Overall, most comments concerned spelling and among them, the majority of comments was about the difference between <-or-> and <-our-> as in examples (14) to (17).

(14) [contributor 1]: good sir I oner you you are the best

[contributor 2]: You mean “honour” or “honor”, British or American language.

While in comment (14), the spelling difference is neutrally explained, in other comments on the <-or->/<-our-> difference, there is a sense of indignation about the use of the American spelling variant. The contributor in (15) outrightly rejects DNO’s use of an American spelling:

(15) NEIGHBOR? Since when do we adopt American English? The word is NEIGHBOUR.

The reader in comment (16) emphasises that the British spelling is the one that “we use” and wonders why DNO uses the American spelling.

(16) [contributor 1]: Look up the meaning between (LABOR and LABOUR) and you will get the difference.

[contributor 2]: Stupes. There is no difference in meaning. Labor is the American spelling and labour is the British spelling which we use. I do not know why DNO is using the American spelling but the story is the same.

Finally, the contributor in (17) is the most explicit in asserting that there is a national norm (‘Dominica English’) from which American spellings deviate.

(17) Could we PLEASE have the DOMINICA ENGLISH spelling of words? Like COLOUR and not “color” as in American English. I have noticed that the English speaking Caribbean countries are very Americanised these days.

A similar stance on forms to be used in Dominica is visible in comment (18), which is about one of the non-systematic differences in spelling. Here <check> is apparently not recognised as the American variant by the contributor and interpreted as a spelling error. After the administrator has explained the matter, another contributor comments further on the issue of American spellings, expressing disapproval and emphasizing that when such spellings are used the words are “not spelt as they are in DA [Dominica]”.

(18) [contributor 1]: Congrats! DNO please do a spell check... how can you spell cheque as (check)?

> ADMIN: American English refers to it as a “check“. British English refers to it as a “cheque“. They are both correct.
>
> [contributor 2]: […] Where have you been and what are you reading? Some years ago Americans have changed some words and/or eliminated some words as also the “u” in behaviour, labour, harbour, etc. Also the letter ‘r’ in centre. They write ‘center’ and some others. I do think they should have left them alone. Whenever you notice those words, not spelt as they are in DA, you will understand.

In comment (19), the possibility of a difference between British and American English is suggested although the case in question is purely a spelling error.

(19) actually love the word is Consensual not consentual the t is dropped in the transformation
Idk if there’s a difference between the English and the american spelling however so you could both be right.

Thus one can see that apart from the major well-known systematic differences, of which <-our> versus <-or> seems to be the most salient, variation between British and American spellings is currently causing some confusion among DNO readers, with American variants sometimes suspected to be spelling errors and vice versa.

Only one single comment about an aspect of lexis in DNO was found and this is again not an actual case where British and American English differ; rather the writer ascribes a usage he or she considers as incorrect to American English:

(20) DNO please use proper titles in your stories: a woman cannot be a Chairman. This is modifies American English which is unacceptable.

Apart from this comment, another two lexical differences between British and American English are mentioned in comments bringing up aspects of American English in the course of a discussion rather than reacting to a specific instance of language use: American *truck* versus British *lorry* (one contributor writes about this in the context of a discussion that is actually rather about whether certain types of vehicles used in Dominica should be properly referred to as *(pickup) truck, van,* or *bus*) and American *soccer* versus British *football* (mentioned in connection with other aspects of American speech and behaviour). Generally, the use of American lexis does not provoke readers to post negative comments.

There are slightly more comments relating to aspects of grammar. In comment (21), the regular verb form *learned* is rejected in favour of the irregular form *learnt*.

(21) And the word is learnt, not learned [...].

This situation is reversed in comment (22), where a contributor considers the irregular verb form *learnt* in the first line of an DNO article ("Dominica News Online has learnt that police are investigating the death of a man who has been found hanging in his house in Woodfordhill.") as a spelling error and the administrator has to explain that it is in fact a British English form.

(22) [contributor]: C'mon people, learn to spell. First of all D.O., there is no such word as learnt. The damn word is learned. [...]

ADMIN: Please do a little research first. Learnt is British English while Learned is American. Last we checked we were a British colony hence we continue using the British way of spelling.

It should be noted in connection with this exception to the general trend of the comments that *learnt* versus *learned* (as past tense or past participle of *learn*) does not represent a clear-cut difference between British and American English in the way contrasts in spelling like <-our> versus <-or> do. *Learnt* is in fact only a minority variant in British English; for example, in the press section of BE06 there are four instances of *learnt* and twice as many of *learned*.[13]

In comment (23), non-standard usage is perceived to be American English:

(23) contributor 1: [...] "there are no evidence" is definitely grammatically wrong as is "I've rode my unicycle", the correct form being "ridden". [...]

contributor 2: [...] Having "rode" her unicycle has its roots in "American English" and those of us who adhere to "Her Majesty's English", can't stand it. Americans rarely use the past participle. Many TV personalities

[13] In the whole BE06 corpus the numbers are 18 and 46, respectively. Cf. also Hundt's (1998: 30) results for large data sets from the *Miami Herald* and the *Guardian* from the early 1990s: the proportions of the variant *learned* in the data from these newspapers are 100% and 78%, respectively.

say, “I had went, “have drove”, “have wrote” etc. Sometimes I wonder, but it’s a common occurrence.

The text whose language use is under discussion here is a commentary and it is mentioned that the author is based in the USA, which may have influenced contributor 2 in making the connection. However, his or her impressions being based on watching television he or she is also likely to be more familiar with American than with British language use as American television is very prominent in the Caribbean nowadays (see e.g. Roberts 2007: 9).[14]

In comment (24), a usage perceived as incorrect is again attributed to American English:

(24) [contributor 1] […] i don’t understand these media mediums on the Island. […]

[contributor 2] […] please, don’t embarrass yourself and our education system; Media = the plural of Medium(too much American grammar mangling creeping in).

Finally, there is also a comment on contractions that describes the common practice in North American English to use contracted forms.

(25) In Canada and the U.S. the norm is to use the word ‘can’t’ rather than ‘cannot’. I also notice this on the Internet when reading the news or other matters. There are also some typographical errors.

Overall, one can see that of the phenomena investigated in the analyses of the corpora in Section 3 only spelling generates controversy. That the grammatical features investigated seem to go largely unnoticed is not surprising given that the differences are only quantitative. However, the view in (25) describing the widespread use of contractions in writing as characteristic of North American English fits in with the corpus findings, which also revealed a sharp difference in the case of contractions. Other comments in the area of grammar revealed some negative attitudes towards American English which are probably at least partly due to the

[14] Specifically in Dominica, the situation is described as follows in the CIA World Factbook (Central Intelligence Agency, 2018): “no terrestrial TV service available; subscription cable TV provider offers some locally produced programming plus channels from the US, Latin America, and the Caribbean”.

fact that there is more exposure to American English, including non-standard varieties, through media, whereas the now more remote British English may be more automatically associated with “Her Majesty’s English”, as in example (23) above. In view of the resistance on the part of some of the contributors to American spellings and what are perceived as American grammatical usages it may be hypothesised that the contributors are not necessarily aware that quite a number of lexical items commonly used in Caribbean press language are actually associated with American English. Further research using an elicitation method along the lines of Sand (2011) would be needed to substantiate this possibility.

5 Discussion

The present paper has investigated the role that American English plays as a factor in Standard English in the press in Jamaica and the small Caribbean island nations of St. Kitts & Nevis, Dominica and St. Vincent & the Grenadines. Analyses of language use as documented in corpora were combined with the perspectives of newspaper staff and readers of press texts.

With regard to spelling the results were mixed. American English turned out to be quite influential in this respect in two of the small island nations, St. Kitts & Nevis and Dominica. In the case of St. Kitts & Nevis about equal proportions of American and British spellings were found in the corpus. The interviews revealed that this reflects a situation where to some extent variation is the norm, as two of four newspapers explicitly accept both British and American spellings. In the Dominican corpus the overall proportion of American spellings is quite high as well. In one of the interviews with Dominican sources it was also stated that both British and American spelling is accepted. The fact that another source said that spelling norms were currently under review, combined with the complaints made by readers in online comments, suggests that this is an aspect that is at present undergoing change. In the corpus data from St. Vincent & the Grenadines the proportion of American spellings is comparatively low and according to the interviews those that occur are rather due to a lack of enforcement of editorial policy than to an actual change in norms. Only a very low proportion of American spellings was found in the Jamaican corpus and the interviews corroborated that the British spelling model remains firmly in place. However, as we have argued, that does not necessarily imply a strong attitudinal orientation towards British English as an exonormative standard, since a local standard, which incorporates British spellings, is recognised in Jamaica. Thus, the fact that more American spellings were found in the data from the small Caribbean nations can be interpreted to show that language use there is more susceptible to (varying) exonormative influences, as has also been proposed by Baker and Pederson (2013: 72) in their work on St. Kitts & Nevis:

> A century ago, one would have expected that, the smaller the territory and its population, the greater would be its uniformity of language. But with exposure through the media and tourism to essentially the same wide range of "outsider" varieties of English, the influence of the latter is potentially far greater on smaller speech communities than large ones.

However, a degree of endonormativity on an attitudinal level was also observed when readers of DNO objected to American spellings on the grounds that they violate what is perceived as the Dominican norm.

Whereas in the case of spelling a considerable degree of correspondence between the corpus data and stated editorial practice was found, this was less the case with lexis. Moreover, no objections to, and few comments on, American lexical items were found in the search on DNO. This may suggest that lexical items such as the admittedly small set that this study has found to be commonly used across the different Caribbean corpora are so established in local usage that they are not salient as American influences in the way e.g. the <-or> spelling is, though further research would be needed to confirm this. Apart from that a common pattern was observed in that the preferences for either the American or, in a small number of cases, the British lexical items are the same in all of the Caribbean corpora.

A common tendency across the different Caribbean corpora was especially apparent for the grammatical features investigated. These all belong to a group of features (see Leech et al. 2009: 271 for a list) that can be related to the trend of colloquialisation that has been manifest in both American and British English, but in the former more than the latter. The analyses have shown that, as measured by these features, the corpus data from the different Caribbean countries are united by a degree of formality that sets them apart not only from American but also from British press language. One case, that of contractions, where the difference in frequency was found to be greatest, even attracted a comment on DNO.

6 Conclusion

Overall one can say, considering the findings on spelling and lexis, that American English is more influential than some may think or like it to be, and British English less. However, the competition between these two different exonormative standards is insufficient to account for all the findings. In particular, the current norm in the English of the press in all the Caribbean countries considered here is to use certain formal grammatical options more, and informal ones less, than in both of the metropolitan standards.

The present research sheds a first light on Standard English as used in small Caribbean countries. This, as has been shown, may have dynamics of its own while also sharing features in common with the Standard English of a relatively large country like Jamaica. Thus, to gain a more complete picture of the current state of Standard English in the Caribbean it is important not to leave small countries out of consideration. Future research could also examine spoken English e.g. as used in broadcasting. Methodologically, the present study has shown that for a fuller appreciation of the issue of Standard English in newspapers the findings gleaned from corpora can be complemented by taking into account the perspectives of newspaper staff and readers. On a more general level this study has highlighted the need for research on Standard English to take both usage and attitudes into consideration.

References

Allsopp, R. (1996). *Dictionary of Caribbean English Usage*. Oxford, UK: Oxford University Press.

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, *14*(3), 312–337.

Baker, P., & Pederson, L. (2013). *Talk of St Kitts and Nevis*. London, UK: Battlebridge.

Bruckmaier, E., & Hackert, S. (2011). Bahamian Standard English: A first approach. *English World-Wide, 32*(2), 174–205.

Bryman, A. (2012). *Social Research Methods* (4th ed.). Oxford, UK: Oxford University Press.

Central Intelligence Agency. (2018). The World Factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook (last accessed October 2018).

Christie, P. (1989). Questions of standards and intra-regional differences in Caribbean examinations. In O. Garcia & R. Otheguy (Eds.), *English across Cultures, Cultures across English: A Reader in Cross-Cultural Communication* (pp. 243–262). Berlin, Germany: Mouton de Gruyter.

Craig, D. (1997). The English of West Indian university students. In E. Schneider (Ed.), *Englishes around the World, vol. II: Caribbean, Africa and Australasia. Studies in Honour of Manfred Görlach* (pp.11–24). Amsterdam, Netherlands: John Benjamins.

Curzan, A. (2014). *Fixing English: Prescriptivism and Language History*. Cambridge, UK: Cambridge University Press.

Deuber, D. (2009a). “The English we speaking”: Morphological and syntactic variation in educated Jamaican speech. *Journal of Pidgin and Creole Languages, 24*(1), 1–52.

Deuber, D. (2009b). Standard English in the secondary school in Trinidad: Problems – properties – prospects. In T. Hoffmann & L. Siebers (Eds.), *World Englishes: Problems, Properties and Prospects* (pp. 83–104). Amsterdam, Netherlands: John Benjamins.

Deuber, D. (2010a). Modal verb usage at the interface of English and a related Creole: A corpus-based study of *can/could* and *will/would* in Trinidadian English. *Journal of English Linguistics, 38*(2), 105–42.

Deuber, D. (2010b). Standard English and situational variation: Sociolinguistic considerations in the compilation of ICE-Trinidad and Tobago. *ICAME Journal, 34*(1), 24–40.

Deuber, D. (2013). Towards endonormative standards of English in the Caribbean: A study of students' beliefs and school curricula. *Language, Culture and Curriculum, 26*(2), 109–27.

Deuber, D. (2014). *English in the Caribbean: Variation, Style and Standards in Jamaica and Trinidad*. Cambridge, UK: Cambridge University Press.

Deuber, D., & Leung, G. (2013). Investigating attitudes towards an emerging standard of English: Evaluations of newscasters' accents in Trinidad. *Multilingua, 32*(3), 289–319.

Garrett, P. (2010). *Attitudes to Language*. Cambridge, UK: Cambridge University Press.

Gut, U. (2011). Relative markers in spoken Standard Jamaican English. In L. Hinrichs & J. Farquharson (Eds.), *Variation in the Caribbean: From Creole Continua to Individual Agency* (pp. 79–104). Amsterdam, Netherlands: John Benjamins.

Hackert, S. (2010). ICE Bahamas: Why and how? *ICAME Journal, 34*(1), 41–53.

Hänsel, E. C., & Deuber, D. (2013). Globalization, postcolonial Englishes, and the English language press in Kenya, Singapore, and Trinidad and Tobago. *World Englishes, 32*(3), 338–357.

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409.

Hinrichs, L. (2006). *Codeswitching on the Web: English and Jamaican Creole in E-mail Communication*. Amsterdam, Netherlands: John Benjamins.

Hinrichs. L., Szmrecsanyi, B., & Bohmann, A. (2015). *Which*-hunting and the Standard English relative clause. *Language, 91*(4), 806–836.

Huddleston, R., & Pullum G.K. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Hundt, M. (1998). *New Zealand English Grammar: Fact or Fiction?* Amsterdam, Netherlands: John Benjamins.

Irvine, A. (1994). Dialect variation in Jamaican English: A study of the phonology of social group marking. *English World-Wide, 15*(1), 55–78.

Irvine, A. (2004). A good command of the English language: Phonological variation in the Jamaican acrolect. *Journal of Pidgin and Creole Languages, 19*(1), 41–76.

Irvine, A. (2008). Contrast and convergence in Standard Jamaican English: The phonological architecture of the standard in an ideologically bidialectal community. *World Englishes, 27*(1), 9–25.

Jantos, S. (2010). Agreement in educated Jamaican English: A corpus-based study of spoken usage in ICE-Jamaica. In A. Wanner & H. Dorgeloh (Eds.), *Approaches to Syntactic Variation and Genre* (pp. 305–331). Berlin, Germany: Mouton de Gruyter.

Kövecses, Z. (2000). *American English: An Introduction*. Peterborough, Canada: Broadview.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge, UK: Cambridge University Press.

Mair, C. (2002). Creolisms in an emerging standard: Written English in Jamaica. *English World-Wide, 23*(1), 31–58.

Mair, C. (2006). *Twentieth-century English: History, Variation and Standardization*. Cambridge, UK: Cambridge University Press.

Mair, C. (2009). Corpus linguistics meets sociolinguistics: Studying educated spoken usage in Jamaica on the basis of the International Corpus of English. In T. Hoffmann & L. Siebers (Eds.), *World Englishes: Problems, Properties and Prospects* (pp. 39–60). Amsterdam, Netherlands: John Benjamins.

Mair, C. (2011). Corpora and the New Englishes: Using the "Corpus of Cyber-Jamaican" to explore research perspectives for the future. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A Taste for Corpora: In Honour of Sylviane Granger* (pp. 209–236). Amsterdam, Netherlands: John Benjamins.

Milroy, J., & Milroy, L. (1999). *Authority in Language: Investigating Standard English* (3rd ed.). London, UK: Routledge.

Oenbring, R. (2013). *Bey* or *bouy*: Orthographic patterns in Bahamian Creole English on the web. *English World-Wide, 34*(3), 305–340.

Peters, P. (2001). Varietal effects: The influence of American English on Australian and British English. In B. Moore (Ed.), *Who's Centric Now? The Present State of Post-Colonial Englishes* (pp. 297–309). Oxford, UK: Oxford University Press.

Rayson, P. & Garside, R. (1998). The CLAWS Web Tagger. *ICAME Journal*, *22*, 121–123.

Roberts, P. A. (2007). *West Indians and Their Language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Rosenfelder, I. (2009). Rhoticity in educated Jamaican English: An analysis of the spoken component of ICE-Jamaica. In T. Hoffmann & L. Siebers (Eds.), *World Englishes: Problems, Properties and Prospects* (pp. 61–82). Amsterdam, Netherlands: John Benjamins.

Sand, A. (1999). *Linguistic Variation in Jamaica: A Corpus-based Study of Radio and Newspaper Usage*. Tübingen, Germany: Gunter Narr.

Sand, A. (2011). Language attitudes and linguistic awareness in Jamaican English. In L. Hinrichs & J. Farquharson (Eds.), *Variation in the Caribbean: From Creole Continua to Individual Agency* (pp. 163–188). Amsterdam, Netherlands: John Benjamins.

Scott, M. (2008). WordSmith Tools version 5. Liverpool: Lexical Analysis Software.

Smith, N., & Leech, G. (2013). Verb structures in twentieth-century British English. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The Verb Phrase in English: Investigating Recent Language Change with Corpora* (pp. 68–98). Cambridge, UK: Cambridge University Press.

Tottie, G. (2002). *An Introduction to American English*. Malden, MA: Blackwell.

Trudgill, P. (1998). World Englishes: Convergence or divergence? In H. Lindquist, S. Klintborg, M. Levin, & M. Estling (Eds.), *The Major Varieties of English: Papers from MAVEN 97*, Växjö 20–22 November 1997 (pp. 29–34). Växjo, Sweden: Växjö University.

Winford, D. (1991). The Caribbean. In J. Cheshire (Ed.), *English around the World: Sociolinguistic Perspectives* (pp. 565–584). Cambridge, UK: Cambridge University Press.

Winford, D. (1993). *Predication in Caribbean English Creoles*. Amsterdam, Netherlands: John Benjamins.

Appendices

Appendix A. URLs of newspapers and news websites (as of the time of data collection)

Dominica News Online	http://dominicanewsonline.com/news/
Dominica Vibes	http://www.dominicavibes.dm/
The Dominican	http://www.thedominican.net/index.html
The Jamaica Gleaner	http://jamaica-gleaner.com/
The Jamaica Observer	http://www.jamaicaobserver.com/
The Labour Spokesman	http://www.labourspokesman.com/
Searchlight	http://searchlight.vc/
SKN Leewards Times	http://www.sknclt.com/
The St Kitts and Nevis Democrat	http://www.skndemocrat.com/
St Kitts and Nevis Observer	http://www.thestkittsnevisobserver.com/
The Sun	http://sundominica.com/
The Vincentian	http://thevincentian.com/
The Western Mirror	http://www.westernmirror.com/

Appendix B. Telephone interviews

Questions

a) Where are the writers at ________________ from? Are there writers who were born and raised elsewhere than in ________________? If yes, where?

b) Have the writers been abroad for a longer time period? Have they worked or studied abroad? If yes, where?

c) Does ________________ have its own style guide? If yes, what does it specify regarding the use of different varieties of English?

d) Do you have regulations as to what type of spelling (e.g. American English, British English) to use at ________________? So for example, should writers spell the word ‘neighbour’ with O-R or with O-U-R?

e) Should writers at ________________ use vocabulary that belongs to one particular English variety (e.g. American English, British English)? For example, should the writers use "toward" as in American English or "towards" as in British English?

f) Do you use spell checkers (e.g. in Microsoft Word)? If yes, which one?
g) Are the writers supposed to stick to a certain English dictionary? If yes, which one?
h) Does ________________ use news agency articles? If yes, what news agency provides these articles?

Responses

	JAM (N=2)	SKN (N=4)	DMA (N=3)	SVG (N=2)
a) Origin of writers	(I) most from Jamaica (II) most from Jamaica; have had one or two from UK/USA	(I) from SKN, elsewhere in the Caribbean, and the USA[a] (II) most from SKN (III) all from SKN (IV) n.i.	(I), (III) all from Dominica (II) most from Dominica	(I) most from SVG, two from Guyana, one from Nigeria (II) most from SVG
b) Stays abroad	(I) some might have lived abroad (II) some have studied abroad	(I), (II) some have studied/ lived in the USA (III) no (IV) n.i.	(I) one has studied in the USA (II) editor has studied in the USA and lived in Jamaica; owner has studied in Jamaica; two contributors (one from Dominica and one from elsewhere in the Caribbean) live in the USA (III) some have been living abroad for a number of years (USA, England, Guadeloupe)	(I) some have studied and worked in Canada/USA (II) two have studied in Mexico; one lives in the USA
c) Style guide	(I) no (II) yes, based on British rules	(I), (II), (IV) no (III) n.i.	(I), (IV) no (II) thinking about designing one (III) n.i.	(I) n.i. (II) yes, based on British rules

	JAM (N=2)	SKN (N=4)	DMA (N=3)	SVG (N=2)
d) Spelling	(I), (II) British	(I) both accepted, also side by side (II) both accepted (III), (IV) British	(I) British (II) question they struggle with; lately big impact of American style as more and more people use it; currently undecided whether to stick to British English or whether to allow American English as well (but won't switch entirely to American English) (III) no, writers are free to choose	(I), (II) British
e) Vocabulary	(I) British, some American words might be used (II) British	[same as for spelling]	(I) n.i. (II), (III) writers are free to choose	(I) do not insist on that but British English mostly used (II) British
f) Spell checkers	(I), (II) yes; British	(I) yes; decision left to each individual writer (II), (III) n.i. (IV) British	(I) yes; British (II) yes (but don't always remember to run it); British (III) yes; British or American	(I) yes; British (II) have Microsoft Word spell checker but usually don't use it
g) Dictionary	(I) writers may use British or American dictionaries but use British spelling either way (II) no	(I), (II) n.i. (III) *Oxford English Dictionary* (IV) Oxford	(I), (II) n.i. (III) no, writers are free to choose	(I) mostly *Concise Oxford Dictionary*, but writers are free to use others (II) proofreaders use *Longman Contemporary Dictionary of English*

	JAM (N=2)	SKN (N=4)	DMA (N=3)	SVG (N=2)
h) News agency articles	(I) yes, press releases, mainly local (II) yes, from AP & Caribbean Media Corporation	(I) n.i. (II) n.i. (III) yes, from Reuters (IV) no	(I) yes, from BBC (II) no, but open to contributions from different writers (III) no	(I) no (II) n.i.

n.i. = no (or unclear) information.

[a] It was unclear to what extent the answer referred to regular writers or other contributors.

Appendix C. Results for spelling by category (raw numbers)

	JAM	SKN	DMA	SVG	AE06	BE06
<-ll-> (e.g. *fulfill*, *willfully*)	2	10	3	7	24	0
<-l-> (e.g. *fulfil*, *wilfully*)	10	6	3	3	0	5
<-l-> (e.g. *traveling*, *fueled*)	7	29	32	11	62	0
<-ll-> (e.g. *travelling*, *fuelled*)	121	38	20	26	1	65
<-ense> (e.g. *license*, *defense*)	6	22	25	6	44	0
<-ence> (e.g. *licence*, *defence*)	63	41	31	36	0	40
<-iz-> (e.g. *realize*, *globalization*)	52	212	268	301	276	1
<-is-> (e.g. *realise*, *globalisation*)	218	136	70	89	0	171
<-yz-> (e.g. *analyze*, *paralyzed*)	0	5	2	6	3	0
<-ys-> (e.g. *analyse*, *paralysed*)	6	2	2	4	0	3
<-ter> (e.g. *meter, center*)	2	23	8	3	98	0
<-tre> (e.g. *metre, centre*)	66	17	10	28	0	93
<-gram> (e.g. *gram*, *program*)[a]	0	122	99	5	145	0
<-gramme> (e.g. gramme, *programme*)[a]	107	27	66	118	0	29
<-or> (e.g. harbor, *labor*)[b]	2	73	56	19	214	0
<-our> (e.g. harbour, *labour*)[b]	193	151	89	151	1	175
miscellaneous AmE: *check(s), cozy, gray, jeweler(s)/-ry, skeptic(s)/-al/-ism, tire(s)*[c]	4	15	9	7	26	0
miscellaneous BrE: *cheque(s), cosy, grey, jeweller(s)/-ry, sceptic(s)/-al/-ism, tyre(s)*	30	8	16	9	0	52

[a] There were no occurrences in the sense of ‘computer program’, where the spelling <-gram> is usually used in British English as well.

[b] Excluding *glamo(u)r* as in this word the <-our> spelling is favoured in American English as well.

[c] *Check(s)* and *tire(s)* were only included when used as nouns.

Alessandra Molino

Corporate identity and its variation over time

A corpus-assisted study of self-presentation strategies in Vodafone's Sustainability Reports

Abstract: This study explores the discursive construction of corporate identity in disclosure statements about non-financial performance. The corpus for analysis consists of the Sustainability Reports published by Vodafone, the British telecommunications company, over twelve fiscal years, from 2000/2001 to 2011/2012. After retrieving instances of self-references in subject position and quantifying them, the collocational profile of the two most frequent forms, i.e. *Vodafone* and *we*, will be described paying attention to the textual meanings most often associated with them through the analysis of concordance lines and their classification in functional groups (Mahlberg 2007). Due to its relatively long and consistent tradition of social and environmental reporting, Vodafone is eligible for a case study of whether and how its corporate identity has changed over time. Therefore, self-presentation patterns will be examined not only in the corpus as a whole, but also in individual subcorpora to gather evidence of possible rhetorical shifts in the way Vodafone has shaped and reshaped its corporate identity.

1 Introduction

In today's socio-economic context, corporations have fully-fledged legal personalities: they can "undertake actions, own property and do business in [their] own name" (Breeze 2013: 4). Companies have become proper social agents, who have to respond for the consequences of their activities to the communities in which they operate. The globalisation of the market together with mounting competition have further increased the need for businesses to publicly account for their activities (Evangelisti Allori & Garzone 2010: 10). Therefore, it is crucial for corporations to maintain good relationships with different groups of primary and secondary stakeholders. Communication is key to the company's success. Much

Alessandra Molino, University of Torino, alessandra.molino@unito.it

https://doi.org/10.1515/9783110489071-004

of it has now to do with the construction of a credible public identity that is capable of generating social consensus on the corporation's activities.

According to Breeze (2013: 8–15), identity is the projection of the company's self-understanding (see Balmer & Greyser 2002 for the view that corporations have not one, but multiple identities that need to be aligned to work effectively). Identity is the result of top-down decisions taken at managerial level and is normally kept under strict control because it is a powerful tool for the company to differentiate itself from competitors. As indicated in *The 1st Strathclyde Statement on Corporate Identity* (International Corporate Identity Group, ICIG, 1995), identity can reinforce organisational culture and ultimately guarantee growth and success.

There is broad consensus among communications scholars and discourse analysts that corporate identity is constructed through discourse. As Breeze (2013: 178) puts it,

> all the company's relationships, with clients or customers, government or state, competitors, investors, stakeholders in general and, of course, the media, can and should be managed through discourse. Discourse is one of the corporation's most powerful tools in the current configuration of society.

As discourse is as a form of social practice performed through the use of language (Jones 2012), by analysing how corporations employ language, it is possible to gain insights into the way they construct their identity reflecting and reinforcing the system of beliefs and knowledge on which they rely. This paper deals with the way corporate identity is discursively constructed in non-financial disclosures, i.e. documents issued by companies embracing the values of sustainable development and reporting on their social and environmental performance. Previous studies (e.g. Aiezza 2015; Lischinsky 2010, 2011) have shown that one of the primary functions of non-financial disclosures is not so much to provide performance data, but to construct for the company the identity of a responsible business. Examining what language choices are made to achieve such a goal will provide insights into the way companies represent themselves in relation to the issue of sustainability and how they propagate certain values. This is, precisely, the overall aim of this study, which will be pursued through three more specific research goals – as explained in the remainder of this introduction – analysing how the British telecommunications company Vodafone talks about itself and its actions in a corpus of Sustainability Reports (SRs) covering twelve fiscal years, from 2000/2001 to 2011/2012.

A first goal will be to understand what type of 'persona' the company projects when talking about its sustainable development policies and practices, specifically whether it considers it to be more rhetorically effective to appear as a formal and institutional entity or as a community with shared values and objectives. To this aim, I will investigate what forms of self-reference are preferred by Vodafone in its SRs, assuming that different types of person deixis imply different kinds of self-understanding. In line with existing corpus-assisted discourse studies of corporate identity (Aiezza 2015; Lischinsky 2010, 2011), I will concentrate on the 1st person plural pronoun *we* and 3rd person references (e.g. the company name and expressions such as *the company* and *the Group*), which will be retrieved and quantified by using Sketch Engine's (Kilgariff et al. 2014) concordance programme.

To further elucidate the discourse practices through which Vodafone constructs its identity of a sustainable business, I will look at the textual meanings related to particular types of self-reference in subject position, thus gaining insights into what the company predicates about its various corporate rhetors. To pursue this second goal, I will first analyse what verbs co-occur with the two prevailing forms of self-reference, namely *Vodafone* and *we*. Then, I will explore recurrent textual meanings associated with particular subject-verb combinations. Specifically, I will analyse the concordance lines in which collocations occur and group together instances with similar 'local textual functions' (Mahlberg 2007) (see Section 3.3 for more details). In so doing, it will be possible to identify the meanings that are relevant to the community that produces and consumes a certain text or group of texts (Mahlberg 2007: 196), in this case, SRs. Therefore, the analysis of collocations and their textual meanings will show what activities contribute to the construction of Vodafone's identity as a sustainable business, and what rhetor-activity combinations are considered most effective to this aim, reflecting the type of organisational culture that the company wishes to reinforce through its discourse practices.

The two research goals stated above, i.e. the analysis of preferred forms of self-reference and of the textual meanings that tend to be associated with them, will provide information on the mechanisms of identity construction most frequently adopted by Vodafone. However, considering that corporate identity is "inherently subject to evolution and change" (Evangelisti Allori & Garzone 2010: 12) and that "reporting social and environmental activities has gathered momentum in the last 15 years" (Breeze 2013: 166), a third goal of this study will be to analyse whether Vodafone's corporate identity has changed in recent years responding to specific socio-cultural and organisational demands, and whether any traces can be found of the trends noted in the existing literature (see Section

2 for an overview) about businesses' changing conceptualisation of their identity in relation to sustainability. More specifically, the analysis will concentrate on how self-presentation strategies and the meanings associated with prevailing forms of self-reference have varied since the early 2000s. The occurrences of self-references through the two most frequently used rhetors, i.e. *Vodafone* and *we*, will be investigated in terms of their ratio over time (from 2000 to 2012). Subsequently, a collocation analysis of the main corporate rhetor employed by the company throughout the years, i.e. the pronoun *we*, will be conducted to understand what textual meanings persist among the most salient ones and whether some meanings emerge or recede at specific points in time.

The discourse-analytical approach taken in this study is primarily text-oriented, although the context of communication will be taken into account for the interpretation of results. Indeed, the case study nature of the investigation will allow me to pay greater attention to non-linguistic factors that could influence the interpersonal choices made by the company. Corpus linguistic techniques will be used to notice recurrent patterns of self-presentation. The size of the corpus being modest (about 350,000 words), this study can be considered a 'small-scale corpus analysis' (Bednarek 2009: 21), in which quantitative information is complemented with the manual annotation of meanings.

The paper is organised as follows. Section 2 introduces the notion of Corporate Social Responsibility, i.e. the framework within which companies articulate their commitment to sustainable development. The section moves on presenting the SR as a genre and reviewing existing discourse studies, including corpus-assisted ones, related to the construction of corporate identity in SRs. Finally, Section 2 outlines how the meanings conveyed by SRs have shifted in recent years as documented in previous analyses. Section 3 illustrates the methods adopted in this study, describing the features of the corpus (3.1) and the procedures followed to retrieve self-references (3.2); to derive collocations and group occurrences in functional groups (3.3); and to identify rhetorical shifts across the years (3.4). Sections 4, 5 and 6 present the results of the three research goals of the study respectively, namely what types of self-reference are favoured by Vodafone; what the major verbal collocates of preferred forms of self-reference are and what textual meanings they convey; and, finally, whether the overall picture obtained about Vodafone's strategies of identity construction applies to all the SRs of the corpus or whether rhetorical shifts can be noticed. Section 7 concludes the paper.

2 Corporate social responsibility, Sustainability Reports and corporate identity

Bhatia and Lung (2006: 273) observe that corporate identity results from the combination of various aspects among which are “the values, mission and philosophy” of the corporation. Due to the pressure of growing public awareness (Bhatia 2011: 27) and the interest of investors in buying “clean investment products” (Selmi-Tolonen 2011: 45), businesses are increasingly articulating their identity in terms of commitment to the values of sustainable development. They do so through the notion of Corporate Social Responsibility (CSR). CRS has become a crucial legitimacy strategy for companies: “[t]he development of CSR programmes can be seen as an attempt to narrow legitimacy gaps and cope with them in a systematic fashion” (Ihlen 2009: 245). The commitment to sustainability through CRS therefore enables corporations to enhance, maintain and repair their reputation.

Activities related to the policy of sustainable development are reported annually in disclosure documents called CSR Reports or Sustainability Reports (this is the label used in this paper in accordance with Vodafone’s predominant reporting practices up to the last fiscal year analysed, i.e. 2011–2012). These non-financial disclosures are issued on a voluntary basis. Whether they become an integral part of business processes depends on a number of internal factors, including business size, industry sector and the type of relationship with customers, as well as external aspects, such as the role of investors, public pressure and political regulations.

The SR is a hybrid genre and reflects the “promotional turn” (Breeze 2013: 180) of corporate discourse, whereby informative and persuasive purposes intermingle. A trend has been noticed towards the increasing standardisation of the topics covered, such as information about stakeholder inclusiveness and the sustainability context. This tendency is connected to the adoption of criteria for sustainability reporting recommended in documents such as the Sustainability Reporting Guidelines issued by the Global Reporting Initiative.[1] However, companies do not have to adhere to specific standards, particularly in terms of *how* they communicate contents (Catenaccio 2011). It is this freedom in drafting disclosures and conceptualising what CSR and sustainable development mean

[1] For up-to-date information about sustainability reporting standards, see the Global Reporting Initiative’s website at https://www.globalreporting.org/information/sustainability-reporting/Pages/default.aspx (last accessed August 2018).

for companies that makes SRs particularly interesting to study from a discourse analytical perspective.

CSR reporting strategies have been widely examined, mainly in disciplines such as business management and policy studies. Comparatively, there have been fewer discourse analyses of sustainability disclosures. However, a growing body of research is emerging that employs both text-oriented qualitative approaches and corpus-assisted methods. Among the text analyses is Fuoli's (2012) study of appraisal resources in BP's and IKEA's 2009 social reports. The two companies appear to have divergent approaches: while BP favours the image of an authoritative but detached interlocutor, IKEA appears more empathetic with the audience. Fuoli explains this difference using legitimacy theory and making reference to the different groups of stakeholders primarily addressed by the two corporations, namely investors and regulators for BP and customers for IKEA. Fuoli's study demonstrates that while trying to convey a desired image for themselves, companies also construct relationships with readers that elicit specific attitudinal responses.

Lischinsky (2011), on the other hand, uses both qualitative and quantitative corpus methods to explore self-references in 50 non-financial reports by large Swedish corporations. He observes that impersonal references to the company name are more frequent than 1st person plural references. However, he argues that both impersonal legitimacy and the affiliative voice associated with the use of the pronoun *we* are needed by corporations, who skilfully shift from one form to the other to reach their discourse goals. Lischinsky (2011) focuses, among other phenomena, on the corporate rhetor in subject position. He shows that different forms display different co-occurring patterns with verbs: the company name occurs with items that relate to business activities while *we* co-occurs with verbs connected with ethical issues.

Subject-verb associations are also investigated by Aiezza (2015). She carries out a corpus-assisted discourse analysis of the construction of corporate identity in SRs published in English between 2008 and 2011 by energy companies operating in the BRIC countries (Brazil, Russia, India and China) and in G8 countries. She focuses on forward-looking verbs in association with *we* and the company name, and analyses semantic prosodies and preferences. The results of her study indicate that projections tend to be optimistic and that their purpose rather seems "to stress a positive behaviour than to provide a full picture of future scenarios" (Aiezza 2015: 74). Hence Aiezza argues that positive forecasts mainly serve to enhance corporate image and to stress that the company is already a successful sustainable business.

This paper too is concerned with the meanings conveyed by subject-verb combinations. By focusing not only on lexical verbs but also on auxiliaries and modals, it will be possible to understand whether the actions described are presented as goals, accomplishments or work-in-progress activities. This issue is interesting to investigate in order to understand how Vodafone positions itself with regard to sustainable development. For instance, if the collocation corporate rhetor + *will* were found to be highly salient in Vodafone's SRs, it would suggest that the company emphasises its commitments, expectations and good intents. Hence, the view of sustainability that would derive is one of a yet-to-be-achieved condition, a goal for the future, not a present state of affairs. Since the way companies talk about their actions reflects their ideas of what being sustainable means, corpus-assisted discourse studies that have analysed the notion of sustainable development as understood by companies are also relevant to the present investigation.

Alexander (2002) adopts a critical perspective on the use and content of the very term *sustainable development* as it occurs in the 1999 and 2000 reports by Shell on the topics of people, planet and profits. Using a concordance programme and deriving the left collocates of the phrase, Alexander notices that *sustainable development* has a very positive semantic prosody serving to emphasise the good intentions of the company. However, he also notices that the adjective *sustainable*, when used with other head nouns, has no clear referential value: vagueness in meaning makes this adjective little more than a buzzword which is used strategically by the company to dissipate critical voices. Similarly, Lischinsky (2010) explores the notion of sustainable development looking for the node word *sustain** in a corpus of reports by 50 large Swedish companies. He notices that issues of profit are strictly connected to the concept of sustainability; he also observes that statements of intentions are more salient than references to concrete actions. Brown (2013) combines corpus analysis (keywords and collocations) and framing theory to identify differences and similarities in the meanings conveyed by NGOs and British 'green' corporations in their websites about environmental protection. He finds that the two groups have distinct cognitive systems of concern for the natural environment; businesses, in particular, rely heavily on meanings related to the frames of good intentions and risk management. The former is absent from the NGOs' cognitive system, suggesting that it is specific to the way corporations articulate their commitment to environmental sustainability.

What these studies seem to agree on is that the notion of sustainable development is an elusive one and that companies frame their relationship with it mainly in terms of commitment to its values, showing willingness to listen to the concerns of stakeholders and to address them responsibly. However, there seems to be little (linguistic) evidence that companies' talk of sustainable development in their disclosures is actually "being walked" (Laine 2005: 409). The analysis reported in this paper will verify whether in the SRs published by Vodafone, the company articulates its adherence to the values of sustainability merely through continuous commitment or whether other strategies to obtain legitimacy can be noticed.

Finally, as this paper also explores rhetorical shifts over time, studies that have investigated non-financial disclosures longitudinally are taken into account. Despite being a relatively new genre, the SR seems to have undergone some developments in the past fifteen years or so. Probably the most evident one pertains to the name of the genre itself. While at the beginning the preferred way to refer to non-financial disclosures was 'Environmental Report', now the most widespread term seems to be 'Sustainability Report' (Catenaccio 2011). This change in terminology reflects the fact that corporations have progressively integrated sustainable development strategies in three directions, namely environmental, economic and social sustainability. These are the 'three pillars' that constitute the so-called Triple Bottom Line approach. Consequently, SRs nowadays include not only information about environmental policies, but also about social and economic issues.

Other changes have been noticed in rhetorical terms. Bowers (2010) observes that at the beginning of the 2000s companies committed to sustainability as a matter of compliance to norms and regulation, whereas in the late 2000s, they started stressing the economic value of sustainable activities. Breeze (2013: 166), too, reports a shift in recent SRs, which differently from the past also include issues related to accountability and transparency.

A diachronic study of how businesses' understanding of the notion of sustainable development has changed over time is carried out by Laine (2010). Using 'interpretative textual analysis', Laine focuses on the non-financial disclosures by three major Finnish companies during the period 1987–2005. He finds that over those two decades the corporations turned what initially appeared as a revolutionary concept into an idea to maintain the status quo. By 2005, the view that sustainability was irreconcilable with the principles of capitalism lost appeal and was replaced by the belief that sustainable development is an attainable goal compatible with prevailing business activities. In addition, while in the 1990s there was evidence of conflicts between corporations, environment and society, by the mid-2000s these contrasts had been mitigated and corporations simply seemed to accept sustainability as a common way of doing business. The focus now is no longer on problems but on the solutions provided by companies while undertaking their ordinary operations (Laine 2010). The analysis of rhetorical shifts over time conducted in this paper will verify whether self-presentation patterns and their verbal collocations (especially auxiliary verbs) can uncover traces of similar, or other, discursive shifts in Vodafone's report archive.

3 Methodology

This section presents the corpus collected for this study, illustrates the procedure followed to investigate self-references, describes the characteristics of the analysis of these items over time and explains how rhetorical changes were identified.

3.1 The corpus

The corpus compiled for this study consists of twelve SRs and totals approximately 350,000 words (see Table 1). The PDF texts were downloaded from Vodafone's report archive, which at the time when this research was carried out listed the reports published from 2000/2001 to 2011/2012.[2] The Sketch Engine (Kilgariff et al. 2014) was used to create the corpus from the PDF files.[3] Once converted to plain texts, all the files were hand-checked for conversion errors, an indispensable step since this study deals with collocation analysis. Stretches of text occurring in tables, pictures and captions were retained because in the most recent SRs multimodal communication is extensively exploited and omitting the textual elements within such visual items would deprive the analysis of important data.

Tab. 1: Corpus for analysis consisting of twelve SRs by Vodafone (2000–2012).

Fiscal year	Words
2000/2001	7,968
2001/2002	12,155
2002/2003	15,794
2003/2004	17,610

[2] The Vodafone report archive is available at https://www.vodafone.com/content/index/about/sustainability/approach-and-reporting/reporting-centre.html (last accessed August 2018).

[3] The Sketch Engine Corpus Query System is available at http://www.sketchengine.co.uk (last accessed August 2018).

Fiscal year	Words
2004/2005	26,195
2005/2006	32,493
2006/2007	37,552
2007/2008	107,164
2008/2009	33,167
2009/2010	36,188
2010/2011	12,597
2011/2012	9,717
TOTAL	**348,600**

The average text length of the SRs collected is about 29,000 words. However, some texts are well below or above this value. For instance, the 2011/2012 SR is shorter, because it is a Summary Report, which was published when the company started using the corporate website as the main tool for disseminating information about its sustainable performance.[4] On the other hand, the 2007/2008 SR is longer. As compared to the previous report, it contains more pages and features less visual material. In addition, it presents a different macro-structural organisation with many more thematic blocks. It is difficult to understand why the 2007/2008 SR relies so much on text. What can be observed, however, is that in various passages the company declares to report information *for the first time* (the phrase occurs 10 times in the 2007/2008 SR and not even once in the previous SR). Because of differences in text length, when comparisons across sub-corpora are made, percentages will be employed focusing not so much on the over/underuse of items, but on (dis)similarities in the way features are distributed within individual texts (see Section 6.1).

3.2 Identifying and quantifying forms of self-reference

In order to achieve the first goal of this study, namely to understand what type of ‘persona’ Vodafone constructs in its SRs, self-references were retrieved using Sketch Engine’s concordance programme. The node words *Vodafone*, *we*, *Group*,

[4] The following year, 2012/2013, Vodafone reintroduced the practice of issuing full SRs in PDF format while preserving the new habit of using the corporate website to provide the entire range of information on sustainability policies. The 2012/2013 SR was not available when this research was carried out.

and *company* were focused on, as suggested by Vodafone itself, which in its SRs declares that "[a]ll references to Vodafone, Vodafone Group, the Group and 'we' [...] mean Vodafone Group Plc and its operating companies" (Vodafone Group Plc, *Corporate Responsibility Report 2004-2005*). Reading of concordance lines was required to distinguish between 'averrals' (Tadros 1993), i.e. statements issued by the company, and comments made by other voices reported in the text. Only the former were analysed and quantified.

3.3 Deriving collocational patterns for *Vodafone* and *we* and identifying functional groups

The second goal of this study is to analyse what Vodafone predicates about its various corporate rhetors, thus gaining insights into what actions are presented by means of an institutional identity and what activities the company seeks legitimation for through a more affiliative 'persona'. In order to reach this goal, collocations were derived for the most frequent forms of self-reference, i.e. *Vodafone* and *we*. The log-likelihood (LL) test was chosen as a confidence-based measure to establish associations. Compared to other tests,[5] LL seemed to be better suited to the analysis of collocation and colligation, as it places emphasis on high frequency lexical verbs, auxiliaries and modals. Verbal collocations were calculated setting the window to the 3rd item on the right of the node word. This span was chosen after comparing the collocations thus obtained with those derived using different spans. With the +3 span, major verbal collocates in different constructions could be identified minimising the risk of excluding important verbs, as it happens with the +1 span (e.g. in <u>*we*</u> <u>*would*</u> *continue*) and with the +2 span (e.g. in <u>*we*</u> *will* <u>*also*</u> *support*). On the other hand, for a wider span the risk was that of counting redundant items, as is the case with the +4 span, e.g. <u>*we*</u> *believe we must* <u>*continue*</u>. The minimum frequency in the +3 span was set at three occurrences to exclude hapax and dis legomena.

As for the cut-off point at which collocates were considered worth analysing, the choice was unavoidably arbitrary. I opted for a ranking system (see Baker 2014: 137), meaning that only the items with the highest LL scores were focused on. This decision is mainly justified on practical grounds, since in some cases the

[5] The Mutual Information (MI) test was also applied to the data. MI is a strength of association test which identifies what is unique in the use of a given item in a corpus. It was discarded in this study for its almost exclusive focus on content words, including very low frequency items, which would have compromised the analysis of colligation patterns.

list of collocates with a minimum frequency requirement of 3 was very long. A cut-off point of twenty verbal collocates was chosen as it was wide enough to allow for variation in meaning and sufficiently manageable for a detailed discussion. When the collocates derived with the set span and minimum frequency requirements did not exceed twenty items, no cut-off point was needed, but the focus was nevertheless on collocations with high LL values.

In order to explore recurrent textual meanings associated with specific subject-verb combinations, the collocations identified for *Vodafone* and *we* (e.g. *Vodafone supports* or *we want*) were analysed in their context of occurrence by reading all the concordance lines in which they appeared. Following the procedure outlined in Mahlberg (2007), concordance lines were grouped according to the 'local textual functions' that subject-verb combinations performed, that is, the meanings that they acquired in specific stretches of text. As Mahlberg (2007: 195) observes, local textual functions represent "the textual components of meanings that are associated with lexical items" and the labels assigned to functional groups are devised *ad hoc* to account for meanings that are "embedded in textual contexts" (Mahlberg 2007: 199). For instance, in her study of the use of the cultural keyword *sustainable development* in newspaper articles, Mahlberg notices that it occurs in 11 main semantic contexts, which include the functional group of 'Conferences and the World Summit', where *sustainable development* is the topic for discussion (e.g. *at the world's biggest summit on* <u>*sustainable development*</u>), or the functional group 'Education in and for sustainable development', where the keyword occurs in contexts making clear references to the educational issues of sustainability (e.g. *and pupils are trying to bring* <u>*sustainable development*</u> *issues into the*). By grouping together instances sharing similar textual functions, it is possible to gain insights into the socially-relevant meanings that items convey in the group of texts under study.

3.4 Characteristics of the analysis of self-references and their collocational patterns over time

The third goal of this paper is to analyse whether the overall identity constructed by Vodafone in the twelve SRs of the corpus characterises the way the company has always portrayed itself from 2000 to 2012 or whether differences can be noticed in individual texts, suggesting that the company has adjusted its discourse practices as a response to specific contextual demands changing its legitimacy strategies.

The analysis of rhetorical shifts across years is based on a corpus that extends over a rather short period of time. Hence, the data obtained provide evidence of ‘brachychronic’ (Mair 1997: 202; also see Gabrielatos et al. 2012: 158) variation. In other words, the time-span is too short to talk about ‘diachronic’ variation in language use, but it is sufficiently spread out to capture rhetorical shifts. Two main reasons account for this time length: on the one hand, the amount of material available is limited, as the first SR ever published by Vodafone covers the fiscal year 2000/2001; on the other hand, the case study nature of this investigation required the creation of a manageable corpus. Despite the limited amount of time covered, the analysis of rhetorical shifts over time was conducted assuming that the dramatic growth of sustainability disclosure in the recent years (Pilot 2011) and the increasing attention of the general public to ethical business management would have had an impact on Vodafone’s self-presentation strategies within the CSR context.

According to Gabrielatos et al. (2012) time-span is not enough as a criterion to describe the features of a diachronic corpus study. Another important aspect is ‘granularity’, which is given by the number of sampling points divided by the time length of the corpus. Gabrielatos et al. (2012: 153) argue that high granularity guarantees high levels of accuracy in the results obtained. The present study is characterised by a low level of granularity (=1), as there are 12 sampling points and the time-span covers 12 years. To increase the granularity, the sampling points would need to be augmented by reducing the interval between them. Unfortunately, this is not possible in the present study, as the number of sampling points is constrained by the specific genre under scrutiny, which is normally published once per fiscal year.

Shifts in the use of language were identified by first comparing the ratio (in percentage points) of the two most frequent forms of self-reference (i.e. *Vodafone* and *we*) across the twelve fiscal years. Subsequently, the collocates of *we*, i.e. the predominant corporate rhetor, were calculated for each SR and a cut-off point of twenty items was chosen. In order to distinguish possible changes in the collocational patterns of *we*, functional groups within each subcorpus were identified following the procedure outlined in Section 3.3. Adjacent years, proceeding chronologically, were compared to notice what functional groups were present in each subcorpus. I did not apply any statistical test to identify significant differences in quantitative terms, because the focus is on the relative salience of functional groups and not on phenomena of over/underuse of specific items.

4 Forms of self-reference

Table 2 illustrates the frequency of the various forms of self-reference found in the corpus, distinguishing between all occurrences and those in subject position in averrals by the company. The most widely used construction in subject position is the 1st person plural pronoun *we* followed – with a considerable gap – by the company name *Vodafone*. This result indicates that overall the corporation seems to prefer a personal type of interaction as a way to gain legitimacy for its operations.

Tab. 2: Forms of self-references (raw frequencies)

Self-references	**All**	**Corporate rhetor in subject position in averrals**
The company	187	23
(The) Vodafone Group (Plc)	1,248	27
The Group	503	31
Vodafone	3,570	790
We	4,661	4,339
TOTAL	**10,169**	**5,210**

The way that companies communicate is likely to reflect their values and type of corporate culture. Eaton and Brown (2002) report that in order to sustain increased competition and to regain the lead in its market sector, Vodafone underwent a cultural change in the mid-late 1990s whereby the company started working to "replace 'command and control' with a 'coaching and collaboration' culture" (Eaton & Brown 2002: 284). Vodafone thus moved from a hierarchical organisational structure to a flat one, where team members are more engaged in decision-making processes. This change is likely to have left a mark on the communication strategies adopted in the company's disclosure documents and the preponderance of *we* seems to confirm this presupposition.

Looking at other corpus-assisted discourse studies of corporate identity through self-references, additional interpretations for Vodafone's interpersonal choice can be proposed. Bernard (2015) found that as compared to Integrated Annual Reports (a hybrid genre disclosing both sustainable and financial performance), SRs issued by the same companies tended to rely more on *we*-references. Although broad generalisations should be avoided, a hypothesis that arises from

Bernard's (2015) study is that a correlation might exist between the SR as a genre and the preference for a more personal interactional style. Hence, one reason for the predominance of *we* in the corpus could be that when ethical business behaviour is the main issue at stake, companies find it rhetorically more effective to use 1st person plural pronouns. *We* enables writers to identify with their arguments, to underscore their contribution to sustainable development and to gain credit for their actions.

Another factor playing a role may be the specific cultural and socio-economic context in which companies operate. Aiezza (2015) found that in SRs by companies in G8 countries, *we* was the most frequent corporate rhetor. Conversely, businesses in the BRIC countries adopted a more "bureaucratic style" (Aiezza 2015: 71) favouring the use of the company name and of the label *the company*. It could therefore be hypothesised that "more mature sustainable practices" (Aiezza 2015: 71) allow businesses to appeal to readers through the voice of an "affiliated speaker" (Cheney & McMillan 1990: 97), representing a community that shares values, interests and goals. This interpretation may hold true for Vodafone: despite operating in both developing and developed countries, its engagement with sustainability issues can be considered mature, as it has been committed to sustainable development for more than 15 years and has been regarded as being capable of setting new standards in sustainability reporting practices (Allison-Hope 2014).

Cultural and socio-economic differences might also explain why the results obtained in this investigation are in conflict with Lischinsky (2011: 268). He found that in a corpus of 50 SRs by some of Sweden's largest companies, 3rd person self-reference by means of the company's name was the most frequent choice with a ratio of 2.05:1 to 1st person plural pronouns. The divergence between Lischinsky's (2011) results and the data obtained here is evident, and even more so if only the frequencies for *Vodafone* and *we* in subject position in averrals are considered. In this case, the ratio is 5.5:1, with the 1st person plural pronoun occurring more than five times as frequently as the company name. It is possible, therefore, that different areas and audiences have dissimilar expectations about what counts as convincing interpersonal practices in SRs.

5 Major verbal collocates of *Vodafone* and *we* and their textual meanings

In the previous section different forms of self-reference were extracted from the corpus. It was noted that Vodafone shows a marked preference for *we* and that it relies predominantly on the company name when an impersonal subject is needed. In this section, the verbal collocates of these two forms will be analysed to understand what representation Vodafone makes of itself through such corporate rhetors. I will start by discussing collocations with lexical verbs, distinguishing between verbs that co-occur uniquely with one node word and those that are shared (even though the grammatical form may be different). I will then proceed by considering verbs such as *has* or *are* that can either be used as auxiliaries or in other constructions. Finally, I will focus on modal verbs.

Table 3 compares the main verbal collocates of *Vodafone* and *we*. Both the co-occurrence counts and the LL values are rather high, especially for the 1st person plural pronoun, so the list of likely collocates is long. Only the first twenty items (ranked by LL scores) are shown in Table 3. The verbs in bold type are unique to the top ranking items of each node, while those in Roman are shared, meaning that they appear, albeit in different word forms, within the first twenty collocates of both items. For instance, the item *acting* (and the lemma ACT) only appear among the twenty top ranking collocations of *Vodafone*, while the lemma HAVE is a shared collocation, which appears in the highest ranking position in both lists, in the form of *has* and *have* for *Vodafone* and *we* respectively. In Table 3, the 'co-occurrence count' indicates the frequency of each collocation (e.g. *Vodafone* with *has*, 247 occurrences), while the 'candidate count' refers to the frequency of each specific word form in the corpus (e.g. *has*, 993 occurrences).

Vodafone co-occurs with items that can be divided into two main functional groups. The first includes verbs related to corporate policies and governance practices; the second comprises verbs referring to public/stakeholder perception of the company and to the recognition of its merits. These groups were identified based on recurrent meanings associated with the top 20 verbal collocates of *Vodafone* within all the concordance lines in which they appeared, following Mahlberg (2007) (see Section 3.3).

Tab. 3: First twenty highest-ranking verbs collocating with *Vodafone*[6] and *we*.

Vodafone				We			
Verb	**Co-occurrence count**	**Candidate count**	**LL**	**Verb**	**Co-occurrence count**	**Candidate count**	**LL**
has	247	993	1,123.255	have	779	1,708	4,815.422
is	355	2,976	1,080.939	are	576	2,537	2,581.650
continues	26	47	168.444	**would**	284	371	2,173.009
will	77	1,111	153.486	will	305	1,111	1,478.079
acting	23	39	153.299	believe	120	183	849.592
does	30	121	134.942	**operate**	101	129	777.475
was	47	426	133.476	**said**	100	150	712.924
seeks	15	21	109.160	**aim**	88	144	602.995
supports	17	36	102.684	continue	89	163	580.305
should	29	223	91.515	do	102	307	534.999
recognises	12	26	71.745	**commissioned**	59	82	435.312
engages	9	12	67.052	recognise	56	71	432.143
worked	14	52	65.497	working	93	401	411.827
participating	13	45	62.889	**want**	52	87	352.421
believes	9	17	57.155	engage	55	131	319.452
acquired	12	43	57.106	**provide**	73	364	300.147
introduced	16	108	54.600	**launched**	52	187	250.807
teamed	6	6	53.649	**plan**	45	119	249.526
joined	8	14	52.552	introduced	43	108	243.901
won	9	21	52.073	**developed**	57	271	239.970

[6] Choosing the +3 span, the collocates listed here regard the node word *Vodafone* as well as country-specific references such as *Vodafone UK* and *Vodafone Spain*.

The first functional group, i.e. that of verbs referring to corporate policies and governance practices, will be analysed considering each individual semantic aspect in turn. The items that describe corporate policies are *seeks, supports, participating, teamed (up)* and *joined*. Each verb has distinct shades of meaning. For instance, *seeks* indicates the efforts in achieving sustainability goals that are related to the telecommunications sector or that require collaboration with other institutional parties (example 1) [bold type and italics mine]. *Joined*, *teamed (up)* and *participating* refer the corporation's involvement in initiatives by charities and other sustainability bodies (example 2). *Supports* underscores Vodafone's help and encouragement to governments and non-profit organisations for their policies (example 3).

(1) **Vodafone** also ***seeks*** to help governments meet their objectives on a number of issues of broader public interest. [2006/2007]
(2) **Vodafone *is participating*** in the Global eSustainability Initiative assessment questionnaire and risk assessment tool (GeSI) [...] [2005/2006]
(3) **Vodafone *supports*** the European Commission's efforts to strengthen security and privacy, and increase industry's obligation to notify consumers of security breaches. [2007/2008].

The item that describes governance practices is *acquired*. Its use is illustrated in example (4). The firm's expansion in China is mentioned as investment made for worldwide development, a choice that is justified, in sustainability terms, as an attempt to address the 'digital divide' issue. Personal references are scattered throughout the passage, but the use of *we* with *acquire* would appear too personal, informal and possibly celebratory in this context.

(4) Our rapid growth has meant that we have inherited a wealth of localbusiness, employment and environmental practices. We have made structural changes to respond to this diversity while, at the same time, establishing the consistency that is proper for a global business. In November 2000, for example, **Vodafone *acquired*** an equity stake of approximately 2.18% in China Mobile (Hong Kong) Ltd [...]. [2000/2001]

The need to maintain a high level of formality is also observable in example 5, which illustrates the second main functional group: public perception and recognition. While the idea of community and teamwork is underlined by the possessive pronoun *our*, which appears in the heading of the passage from where example (5) is taken (i.e. *Been recognised for our transparent communication on tax*), the use of *Vodafone* with *won* conveys a more detached and official tone to the statement. In so doing, the company manages to construct a convincing ethos, promoting itself without offending readers for “indulg[ing] in large amounts of self-praise” (Breeze 2013: 103).

(5) Been recognised for our transparent communication on tax.
Vodafone won the ACCA award for best CR report in 2006 and for tax and public policy reporting in 2007. [2007/2008]

The functional group of public perception and recognition also includes the verb *acting*. Although semantically it does not refer to the fact of being perceived or assessed, it was included in this group because it occurs in passages referring to the opinion that stakeholders have of the company. The 23 concordance lines of *acting*, distributed across most fiscal years since 2003/2004, all feature a sequence that recurs with minor changes: [...] *stakeholder opinion on how responsibly* ***Vodafone is acting*** *regarding mobile phones, masts and health*. The verbs preceding such a long object vary and include *survey, improve* and *report on*. On the one hand, this collocation unveils the formulaic nature of SRs, which was already noted by Alexander (2002: 242) through his analysis of the type-token-ratio of two reports by Shell; on the other hand, it illustrates that the company name can work as an impersonalising device. As Breeze (2013: 160) points out, impersonalisation occurs when there is the risk of negative outcomes. In the sequence above, the 3rd person reference reduces the potential threat posed to the “human face” (Breeze 2013: 159) of the organisation by stakeholder opinion on the most material issues for Vodafone.

With regard to the collocates of *we,* three functional groups were identified that do not seem to be salient with *Vodafone* as subject. The first presents forward-looking statements, and includes the verbs *plan*, *aim* and some uses of *want* (examples 6–8). The second deals with the company’s CSR activities and comprises the following items: *commissioned, launched, developed* and some uses of *provide* (examples 9 and 10). Differently from the institutional and impersonal tone conveyed by the company name, *we*-references enable the firm to construct itself as a responsible corporate citizen who embraces the values of sustainable development with honest intentions. The use of *we* also allows the corporation to underscore its proactive attitude and to take the merits for the positive sustainable outcomes of its CSR initiatives.

(6) **We *plan*** to review our new commitments to take into account new issues that may arise in the future and ensure we meet our stakeholders’ expectations. [2003/2004]

(7) **We *aim*** to recycle 95% of network waste across the Group [...]. [2009/2010]

(8) **We** ***want*** to support adaptations of technology to meet the needs of those with physical and mental disabilities. [2001/2002]
(9) […] **we** ***launched*** a dedicated intranet site on climate change in 2007 […].[2007/2008]
(10) During 2004, **we** ***commissioned*** four research studies into the socio-economic impacts of mobile phones in Africa. [2004/2005]

The third functional group in which *we* occurs deals with metadiscourse meanings, and includes some uses of *provide*, *want* and *said*. In examples (11) and (12), *we* makes it explicit that the reporter's voice coincides with that of the corporate citizen, who is therefore responsible not only for sustainable performance but also for its proper disclosure. In example 12, *we* also contributes to the construction of reliability: through a personal tone, the company expresses empathy for customers' worries about identity theft, stressing that it can be trusted to take care of their personal information.

(11) In the following pages **we** ***provide*** a brief overview of the steps taken over the last year to establish our approach to sustainability in Ghana and Qatar (see pages 12–13) [2009/2010]
(12) Identity theft is a growing problem. **We** ***want*** to reassure our customers that their security is protected. [2007/2008]

Vodafone's SRs are sequential texts published constantly since the fiscal year 2000/2001. As will be discussed below, Vodafone establishes intertextual references between SRs through the pattern *we said we would*+infinitive and *we have*+past participle. The function of this structure is to show readers that the sustainability claims made in previous SRs are followed by concrete actions. The collocation *we said* can therefore be viewed as conveying metadiscourse meanings that go beyond the ongoing discourse and refer to previous events in the chronological sequence (Mauranen 2012 and Ädel 2006 provide a similar interpretation for intertextual references in sequences of university speech-events and in writing respectively). In showing that past forward-looking statements and commitments are followed by actions, Vodafone seeks to gain credibility and accountability. Often these sequences are used to narrate what Dryzek (2013: 159) calls "success stories", where achievements are highlighted, as in example (13).

(13) **We** ***said*** **we** ***would increase*** the number of phones collected for reuse and recycling by 50% by March 2007 (from the 2004/05 baseline) […]. **We** ***have achieved*** our target of increasing by 50% the number of phones collected for reuse and recycling from the 2004/05 baseline. [2006/2007]

Certain high-ranking collocates of *Vodafone* and *we* are shared. These can be divided into the following functional groups: commitments (CONTINUE), corporate stance (BELIEVE, RECOGNISE) and CSR activities (ENGAGE, WORK, INTRODUCE). Let us look at some of these verbs to understand why one subject is preferred to the other in specific contexts. The need to usefully step into and out of the discourse, as noted above, is clearly one reason why personal or impersonal forms are employed. This purpose is particularly evident with the cognitive verb BELIEVE (examples 14 and 15).

(14) **Vodafone *believes*** research is best conducted by independent experts and has established a funding framework for national, regional and international research programmes. [2001/2002]

(15) By making effective use of our resources **we *believe* we *can add*** value, ***develop*** new opportunities and, ultimately, ***make*** a positive difference to the World Around Us. [2001/2002]

When comparing examples 14 and 15, it appears that the personal pronoun is preferred when the corporation makes more emphatic statements about corporate values, almost sounding like slogans ("World Around Us" was indeed the title and the catchphrase of the 2001/2002 SR). However, in some cases there is no apparent discursive reason why one form is used instead of the other. This is the case, for instance, of RECOGNISE. This verb tends to appear in Problem-Solution textual patterns: first the company acknowledges the existence of a Problem related to its business sector or its operations; then it presents its practices and policies as the Solution, thus comforting readers (example 16). This behaviour was also noted in Aiezza (2015: 74) and may be seen as linguistic evidence of what Dryzek (2013) calls the reassuring rhetoric of sustainability discourse.

(16) [...] **Vodafone *recognises*** that there is public concern about the safety of RF fields from mobile phones and base stations. We are committed to showing leadership by making objective information widely available and engaging openly in dialogue with our stakeholders. [2006/2007]

(17) **We *recognise*** that the net carbon benefits of some of these options remain controversial and we continued to engage with stakeholders in 2010/11 to assess the most credible options. [2010/2011]

Impersonalisation does not seem to be the main reason why *Vodafone* is used in those contexts. Indeed, the company name is not the most recurrent choice with the verb RECOGNISE (see Table 3), although the discussion revolves around sensitive legitimacy issues. The company seems to prefer the adoption of *we* as an overall strategy, because in this way they can show greater commitment and create a more personal dialogue with readers as in example (17). Therefore, the choice of the company name as subject is difficult to explain interpersonally and may be due to textual or stylistic reasons.

I will now turn to those verbs that may work as both lexical items and function words. I will focus particularly on their uses as auxiliaries. The main goal of this part of the analysis is to illustrate the type of grammatical meanings (i.e. tense and aspect) associated with *Vodafone* and *we*. I will also include modals in the discussion. For both the company name and the plural pronoun, the verbs HAVE and BE are the highest-ranking collocates. Although LL values regard the full range of structures encompassed by these forms (e.g. possessive *have* and copular uses of *to be*), in both cases, most occurrences feature the verb in its role as auxiliary, respectively in the present perfect and the present continuous forms. For instance, 81% of the concordance lines obtained for *we have* feature present perfect constructions and 75% of the occurrences of *we are* are in the present continuous form. Therefore, the colligational profile of *Vodafone* and *we* suggests that the company places considerable emphasis on undertaken actions and achievements (e.g. *we have already taken practical steps* [2001/2002]; *We have fully achieved 16 of our 21 commitments* [2004/2005]). Equally, the firm stresses the ongoing efforts to be a sustainable business (e.g. *In our own operations, we are using smart metering to help us improve energy management across our network as part of our efforts to manage our carbon footprint* [2011/2012]). This is not to say that forward-looking meanings, promises or estimates are not present. The collocation with the modal *will* is among the highest-ranking ones and verbs such as *aim*, *plan* and *want* are indeed very salient. However, considering the frequent use of auxiliaries and the form of various high-ranking lexical verbs (e.g. *participating, working, commissioned, launched*), it can be argued that Vodafone's legitimation endeavours are expressed not only in terms of commitment to the values of sustainable development but also, and possibly more, in terms of work-in-progress activities and accomplishments.

6 Self-references across years

This section illustrates the data obtained for the distribution of self-references across the years. Subsequently, it focuses on the collocates of *we* and explores whether and how the meanings associated with this pronoun have changed over time.

6.1 Frequencies of *Vodafone* and *we* across years

Considering the marked imbalance between the frequency of 1st person plural references and other forms of self-reference (see Table 2 above), one would expect the use of *we* to be a constant trait of Vodafone's communication strategy. In other words, *we* is likely to be the predominant choice throughout the whole corpus. However, it is possible that the earliest SRs display distinct features, maybe reminiscent of the traditional managerial culture. In order to explore this issue, the distribution of self-references across the twelve SRs of the corpus was analysed.

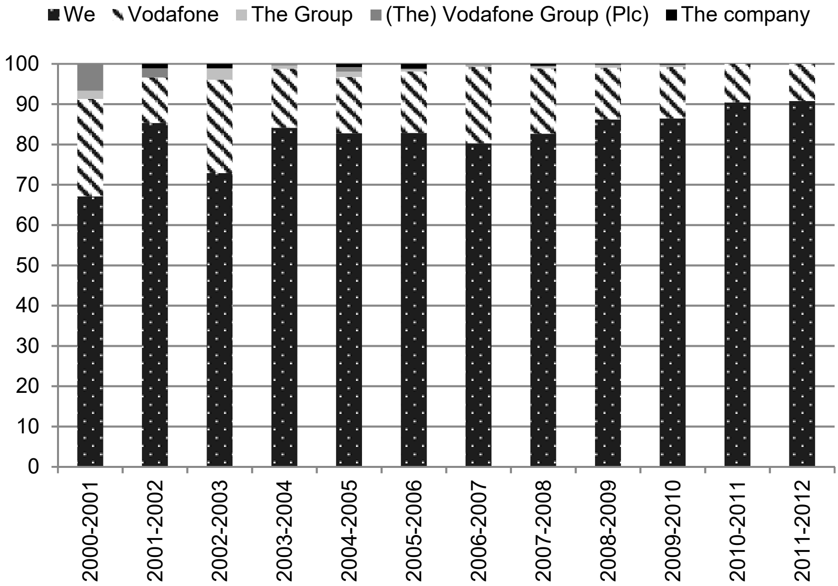


Fig. 1: Self-references across years (percentages).

Figure 1 shows a stacked column chart reporting the results in percentage points. Percentages indicate what portion of the total number of self-references is related to the specific forms. With the exception of the earliest SRs, where indeed a slightly higher proportion of impersonal forms is present, Vodafone's overall construction of corporate identity relies almost exclusively on the 1st person plural pronoun *we* and, to a much lesser extent, on 3rd person references to the company's name. The difference in percentage points between *Vodafone* and *we* increases with the passing of time: the narrowest gap is found in the earliest SR (42.95 percentage points), while the widest one (81.50 percentage points) is observable in the latest document, the 2011/2012 SR. The texts in the middle show a rather constant difference between *we* and *Vodafone* which ranges from 60 to 70 percentage points. By moving towards a more marked preference for an affiliative corporate persona, at least as revealed by the ratio between *Vodafone* and *we*, the company seems to have progressively adjusted its language choices to the changes in policy identified by Eaton and Brown (2002), according to which in the second half of the 1990s Vodafone started working towards a flat organisational structure emphasising teamwork and a sense of belonging to the company. The increasing reliance on *we* as the main corporate rhetor also suggests that this interpersonal choice has proved an effective legitimacy strategy for the company. Whether this feature is typical of the business sector of telecommunications is not possible to establish here, but it would be interesting to understand to what extent this type of 'persona' allows Vodafone to meet the expectations of stakeholders in its sector and to what degree this choice makes the company stand out from its direct competitors. In the following section, I will concentrate on *we* and investigate whether any changes are noticeable in the distribution of the meanings associated with this subject throughout the corpus.

6.2 *We* and its collocates across years

Despite the constant inclination for the use of *we*, a closer look at the collocational profile of this pronoun across years may shed some light on possible changes in the way Vodafone's identity has been represented within the context of corporate sustainability discourse. To recapitulate, the functional groups identified in association with *we* in the corpus as a whole include forward-looking statements (e.g. PLAN); commitments (e.g. CONTINUE); corporate stance (e.g. BELIEVE); CSR activities (e.g. ENGAGE) and metadiscourse (e.g. SAY). Table 4 shows the collocations of *we* across the first six SRs and Table 5 presents the data for the latest six.

Tab. 4: Collocations of *we* (2000/2001–2005/2006).

2000-2001		2001-2002		2002-2003		2003-2004		2004-2005		2005-2006	
Verb	**LL**	**Verb**	**LL**	**Verb**	**LL**	**Verb**	**LL**	**Verb**	**LL**	**Verb**	**LL**
will	127.488	will	185.482	have	168.345	have	405.449	have	404.833	have	463.107
operate	96.815	committed	167.492	are	161.483	are	138.337	would	193.615	would	375.520
are	89.342	have	145.829	committed	78.301	would	115.846	are	166.581	are	233.747
have	87.789	are	139.739	will	68.074	said	95.207	said	116.904	believe	125.643
believe	37.524	believe	90.967	estimate	48.741	also	74.255	will	100.183	said	96.085
intend	31.337	want	84.452	made	46.731	will	72.271	commissioned	61.771	will	95.108
recognise	25.099	operate	46.286	accept	29.167	carried	61.946	aim	61.251	do	61.157
ensure	21.777	determined	30.411	do	29.157	plan	50.973	do	54.975	continue	52.026
continue	20.084	can	30.379	expect	26.859	working	40.974	believe	40.662	operate	48.108
be	13.426	introduce	24.404	recognise	23.177	repeat	35.005	operate	37.189	engage	47.858
can	10.601	made	21.292	encourage	19.763	recognise	28.728	recognise	35.191	provide	44.009
		support	21.100	operate	19.575	need	28.230	met	30.637	developed	42.925
		make	20.783	believe	17.492	track	26.804	engage	28.994	held	42.191
		promote	19.858	set	14.412	complete	26.539	learned	28.707	recognise	39.852
		encourage	19.092	provide	13.462	developed	24.786	set	26.827	aim	38.647
		value	16.481	been	11.204	set	23.507	continue	26.821	conducted	37.075
		developed	14.519			doing	22.567	review	25.310	buy	31.935
		do	13.794			run	20.909	accept	23.938	manage	30.243
		were	12.608			collect	20.706	addressing	23.488	need	27.929
		be	10.249			develop	17.853	explain	22.871	’re	27.524

Tab. 5: Collocations of *we* (2006/2007–2011/2012).

2006-2007		2007-2008		2008-2009		2009-2010		2010-2011		2011-2012	
Verb	**LL**	**Verb**	**LL**	**Verb**	**LL**	**Verb**	**LL**	**Verb**	**LL**	**Verb**	**LL**
have	459.479	have	1,046.766	would	517.182	have	507.270	are	260.314	are	311.620
would	372.673	are	584.499	have	430.067	will	184.199	have	187.844	have	158.083
are	235.554	will	399.425	operate	119.023	would	154.494	commissioned	77.978	working	91.823
will	123.344	would	311.999	are	94.223	aim	143.720	aim	67.460	exploring	75.942
continue	113.804	believe	215.312	do	90.472	are	134.425	continue	55.991	continue	57.748
operate	74.912	want	159.896	said	86.893	believe	98.419	can	55.097	need	55.180
commissioned	65.982	operate	140.684	continue	76.162	said	86.740	expand	46.677	manage	49.135
intend	65.052	said	133.578	will	71.718	working	67.794	face	46.677	work	48.226
said	63.034	provide	126.130	aim	70.626	launched	64.649	believe	37.070	aim	41.267
introduced	60.827	commissioned	125.017	launch	64.741	do	60.739	operate	32.549	will	38.841
believe	59.004	offer	106.459	developed	56.606	commissioned	55.264	recognise	29.051	believe	38.375
held	51.520	aim	103.170	work	44.952	introduced	50.721	continued	28.391	face	36.290
plan	50.530	working	101.837	ensure	41.430	operate	47.827	working	27.793	do	32.210
offer	48.844	continue	98.531	plan	41.168	held	41.946	conducted	27.217	can	26.253
do	48.700	do	94.609	engage	39.813	plan	41.087	targeting	26.484	operate	25.101
provide	45.366	introduced	89.260	provide	37.550	recognise	39.641	exploring	26.235	using	25.000
introduce	44.581	engage	83.199	conducted	36.934	established	38.020	held	25.366	know	24.191
decided	44.005	launched	79.206	did	36.406	asked	37.379	expanding	25.125	achieve	18.437
launched	43.244	consult	77.513	worked	34.419	continue	36.673	made	24.759	must	15.821
engage	40.943	recognise	76.407	revised	32.195	extend	32.217	set	23.902	opportunities	14.811

The functional group of forward-looking statements is almost always present across the years. This indicates that stating intentions and making estimates about future performance are core discourse practices in Vodafone's SRs. The centrality of these meanings is related to the conceptual system that characterises corporations' understanding of sustainability (cf. Brown 2013). As Aiezza (2015: 69) observes, "the function of 'anticipation' is considered a necessary competence to be developed in a CSR report in order to provide a rounded picture of the company's health".

Verbs expressing (or negotiating) corporate stance are also a constant trait. The two most typical collocations of *we* to express intersubjective positioning are *believe* (in the pattern *we believe that*) and *recognise*. These items, however, are not equally salient in all the SRs analysed. In the 2008/2009 document, for instance, *believe* does not appear at all within the twenty highest-ranking items, while it is a collocate of *we* in the 2001/2002, 2005/2006 and 2007/2008 SRs. Such fluctuations suggest that meanings in SRs may be backgrounded or foregrounded to suit to the persuasive needs of the company at specific points in time.

A particularly evident example of the flexibility of the SR as a genre can be found comparing the 2002/2003 and the 2003/2004 SRs. The fiscal year 2003/2004 is a watershed in Vodafone's reporting practices. The company started using the pattern *we said, we have, we will* to underscore its progress against targets (example 18). This tripartite structure shows that claims are followed by actions, thus contributing to the construction of accountability and trustworthiness. Intertextual metadiscourse through the collocation *we said* becomes one of the most prominent meanings conveyed by the 1st person plural pronoun, and it remains so for seven years, between 2003/2004 and 2009/2010.

(18) Last year **we said we would** increase the number of handsets returned for recycling by 10% by 2005 from the 2002/03 level. [...] This year **we have** collected 1.5 million phones. This is an increase of 29% compared to last year's levels, meeting our commitment one year early. [...] Next year **we will** pilot a programme to support reuse and recycling of mobile phones in one developing country. [2003/2004]

The functional group of CSR activities and policies becomes salient for the first time in the 2001/2002 SR. It is expressed though verbs such as *support, promote* and *encourage* and refers to the company's assistance to stakeholders in implementing CSR policies or to its endorsement to third-party initiatives. However, it is interesting to notice that it is in the 2003/2004 SR that Vodafone started emphasising its own proactive behaviour (e.g. *we carried out a major survey of atti-*

tudes towards Vodafone [2003/2004]). This feature will characterise all the subsequent SRs particularly through the verb *commissioned* (see fiscal years 2004/2005, 2006/2007, 2007/2008, 2009/2010 and 2010/2011). Among the company's CSR policies is stakeholder engagement, which seems a particularly noticeable concern between 2003/2004 and 2008/2009.

Finally, the functional group of commitments emerges as salient through the verb *continue,* especially in the mid-2000s. An explanation is that *continue* presupposes previous goals and commitments. Hence, it is a verb that is likely to be found when companies have already made attempts to implement CSR policies. According to Catenaccio (2012: 129–130), current CSR discourse is characterised by an emphasis on ongoing practices. This view is supported by Aiezza (2015), who found that *continue* was the most frequent co-occurring item of *we* in her corpus of SRs published between 2007 and 2011. Considering the timeframe of Aiezza's (2015) study and the data obtained here, it can be hypothesised that the 'work-in-progress' rhetoric of SRs is particularly typical of the second half of the years 2000s, when companies have overcome the scepticism of the 1990s (Schlichting 2013) and fully embraced the values of sustainable development.

If auxiliaries and modals are taken into account, a similar picture emerges. Indeed, a progressive move away from the almost exclusive reference to forward-looking statements and commitments to an emphasis on accomplishments and ongoing activities is noticeable. Evidence for this shift can be found in the ranking of the colligations *we will* and *we are*[7] in 2000/2001 and 2001/2002. It appears that at the beginning of the years 2000s Vodafone's rhetorical strategy was to declare its adherence to the values of sustainable development, which remained, however, largely vague (e.g. *consistent standards*; *to make a real difference*; *to send out a clear environmental message across the world*). The impression is that Vodafone was just starting to experiment with possible ways to make sustainability policies an integral part of its business operations. Evidence for the introduction of new communicative strategies can be found in the ranking of the collocation *we have* starting from 2002/2003 and in the progressive emphasis on ongoing activities that can be noticed in the patterns involving *we are*. In the 2002/2003 SR, 37% (11 raw hits) of the instances of *we are* occur in present progressive forms, while in the 2011/2012 document, the instances of present pro-

[7] In these reports, *are* is most often used as a copular verb with the adjectives *committed* and *determined*.

gressive constitute 81% (44 raw hits) of all occurrences.[8] In the 2003/2004 SR the modal *would* also appears among the highest-ranking items. It signals the company's attempt to gain legitimacy through a rhetoric based on intertextual 'dialogue' between subsequent SRs. On many occasions, progress against a target is documented with recourse to numbers. As Breeze (2013: 184) observes, "'[l]oci of quantity' are a classic means by which a rhetor can intensify his arguments".

7 Conclusion

This study has focused on the discursive construction of corporate identity in a corpus consisting of twelve SRs published by Vodafone from 2000/2001 to 2011/2012. The first goal was to identify the overall type of corporate identity that the company constructed for itself over the twelve years under scrutiny. Of the various forms of self-reference searched in the corpus, the most frequent ones in subject position are the 1st person plural pronoun *we* and the company name *Vodafone*, with a ratio of 5.5:1. The preponderance of *we*, which is a constant feature across the years, was found to be a distinguishing trait of this company, which seems to set Vodafone's rhetorical practices apart from the more common use of the firm name noted in Lischinsky (2011) and corroborated in other studies, such as Swales and Rogers (1995). Vodafone aims to create the image of a dependable corporate citizen willing to engage directly with its stakeholders, taking the responsibility for its actions and emphasizing the cooperative ethical effort of the Group rather than seeking impersonal legitimacy. Explanations for this result may be found in the socio-economic context in which the company operates, but probably the specificity of the telecommunications sector, too, plays a role, as it involves a close relationship with customers who are among the company's major stakeholders. In order to ascertain the influence of business sector, a carefully designed and balanced corpus containing data from firms operating in different areas of activity would be needed. Therefore, this issue remains to be verified in future investigations.

This study has also explored how Vodafone's identity is constructed in relation to what is stated about the two main corporate rhetors employed, which seem to be used for rather different textual functions. The meanings that revolve

[8] The difference is statistically significant according to the Log-likelihood ratio statistics (G^2 = 40.44, $p < 0.0001$; critical value = 15.13), at http://ucrel.lancs.ac.uk/llwizard.html (last accessed October 2018).

around *Vodafone* and its unique lexical collocations refer to corporate policies and government practices, on the one hand, and to the public perception of the company and the formal recognition of its merits, on the other. The impersonal subject in these functional groups serves the purpose of achieving a high level of formality when talking about official relationships with institutional partners, such as governments and non-profit organisations, also enabling the company to maintain a neutral and institutional tone when it comes to acquisitions and awards. The mitigating effect of impersonality can be noticed when Vodafone deals with stakeholder perception about the firm, which in some cases poses a threat to the human face of the corporation.

The textual meanings associated with *we* and its exclusive lexical collocations emphasise the good intentions of the company (i.e. the functional group of forward-looking statements), helping project the identity of a responsible corporate citizen who proactively undertakes actions towards the goal of sustainable development (i.e. the group of CSR activities). A third functional group was identified for *we* and its collocates *provide*, *want* and *said*, that is, the group of meta-discourse meanings. This group is particularly interesting, as it allows Vodafone to align the identity of a responsible corporate citizen with that of a dependable reporter, a choice whereby the company openly takes responsibility for the disclosure of its sustainability performance, thus building accountability and creating a direct relationship with the readers of its SRs. Finally, some collocations of *Vodafone* and *we* were found to be shared and these convey meanings related to commitments, CSR activities and corporate stance. In some cases, it seems that alternation of the two forms enables the company to skilfully manage the level of personality required, but in other cases the choice of one form rather than the other seems to respond to textual or stylistic needs rather than interpersonal ones.

Vodafone and *we* were also analysed in terms of their colligational profiles. Both subjects are used extensively in present perfect constructions (with HAVE) and with present continuous ones (with BE). This result suggests that Vodafone places emphasis not only on goals for the future, although this practice is very salient too, but also on accomplishments and ongoing activities. It was suggested that this strategy is symptomatic of a corporation that has gained experience in devising and implementing policies in favour of sustainable development. The analysis of individual SRs seems to corroborate this hypothesis.

While no evident change was noted in Vodafone's preferred voice over time, some shifts occurred in terms of the collocational and colligational profiles of pronominal references. It is possible to trace a progressive transition of the company's public discourse on sustainability from a marked focus on goals and commitments to rather vaguely defined sustainability goals in 2000/2001, to increased emphasis on the company's achievements and ongoing practices. In addition, starting from 2003/2004 the company strives to project the image of an accountable firm through the *we said, we have, we will* rhetorical pattern. This way of organising the company's sustainability discourse proves an effective strategy that performs simultaneously a variety of functions: it constructs credibility via intertextual references, it gives the idea that sustainability is not a far away and vague target, but a set of very specific short- and medium-term objectives, and it allows the company to restate its commitment to sustainable development portraying it as an ongoing process which also includes not only the future but also the past and the present.

Vodafone's portrayal of its relationship to sustainability has evolved in a way that partly corroborates Dryzek's (2013) description of sustainability as a discourse needing continuous commitment, but which also partly diverges from it, in that it contains features that go beyond promises and that pertain to accomplishments. Whether these achievements have a significant impact in terms of environmental, social and economic sustainability, however, is something that cannot be verified using corpus-based discourse analytical techniques.

This study can be expanded in various ways. It would be interesting to see how self-references are used in specific sections of SRs, thus carrying out more fine-grained analyses of CSR discourse in relation to the environmental, social and economic sustainability. The study of other phenomena, such as longer phraseological patterns, could provide further insights into rhetorical shifts. Finally, replicating this investigation on larger corpora will make it possible to identify more general patterns of the way corporations shape and reshape their identity in relation to the notion of sustainable development.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam, Netherlands: John Benjamins.

Aiezza, M. C. (2015). "We may face the risks"… "risks that could adversely affect our face." A corpus-assisted discourse analysis of modality markers in CSR reports. *Studies in Communication Sciences*, *15*(1), 68–76.

Alexander, R. J. (2002). Everyone is talking about "sustainable development". Can they all mean the same things? Computer discourse analysis of ecological texts. In A. Fill, H. Penz, & W. Trampe (Eds.), *Colourful Green Ideas. Papers from the Conference 30 Years of Language and Ecology, Graz, 2000, and the Symposium Sprache und Ökologie, Passau, 2001* (pp. 239–254). Bern, Switzerland: Peter Lang.

Allison-Hope, D. (2014). What Vodafone's new law-enforcement report says about the future of sustainability reporting. Retrieved from http://www.bsr.org/en/our-insights/blog-view/what-vodafones-new-law-enforcement-report-says-about-the-future-of-sustaina (last accessed October 2018).

Baker, P. (2014). *Using Corpora to Analyze Gender*. London, UK: Bloomsbury.

Balmer, J. M. T., & Greyser, S. A. (2002). Managing the multiple identities of the corporation. *California Management Review*, *44*(3), 72—86.

Bednarek, M. (2009). Corpora and discourse: A three-pronged approach to analyzing linguistic data. In M. Haugh, K. Burridge, J. Mulder, & P. Peters (Eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus* (pp. 19–24). Somerville, MA: Cascadilla Proceedings Project.

Bernard, T. (2015). *A Critical Analysis of Corporate Reports that Articulate Corporate Social Responsibility*. Doctoral dissertation. Stellenbosch University, Stellenbosch, South Africa. Retrieved from http://scholar.sun.ac.za/handle/10019.1/96672 (last accessed October 2018).

Bhatia, V. K. (2011), Contested identities in corporate disclosure documents. In V. Bhatia & P. Evangelisti Allori (Eds.), *Discourse and Identity in the Professions* (pp. 27–44). Bern, Switzerland: Peter Lang.

Bhatia, V. K., & Lung, J. (2006). Corporate identity and generic integrity in business discourse. In J. C. Palmer Silveira (Ed.), *English for International and Intercultural Business Communication* (pp. 265–288). Bern, Switzerland: Peter Lang.

Bowers, T. (2010). From image to economic value: A genre analysis of sustainability reporting. *Corporate Communications: An International Journal, 15*(3), 249–262.

Breeze, R. (2013). *Corporate Discourse*. London, UK: Bloomsbury.

Brown, M. (2013). A methodology for mapping meanings in text-based sustainability communication. *Sustainability*, *5*(6), 2457–2479.

Catenaccio, P. (2011). Social and environmental reports. A diachronic perspective on an emerging genre. In G. Garzone & M. Gotti (Eds.), *Discourse, Communication and the Enterprise. Genres and Trends* (pp. 169–192). Bern, Switzerland: Peter Lang.

Catenaccio, P. (2012). *Understanding CSR discourse. Insights from linguistics and discourse analysis*. Milan, Italy: Arcipelago.

Cheney, G., & McMillan, J. J. (1990). Organizational rhetoric and the practice of criticism. *Journal of Applied Communication Research*, *18*(2), 93–114.

Dryzek, J. S. (2013). *The Politics of the Earth: Environmental Discourses* [3rd edition]. Oxford, UK: Oxford University Press.

Eaton, J., & Brown, D. (2002). Coaching for a change with Vodafone. *Career Development International*, *7*(5), 284–287.

Evangelisti Allori, P., & Garzone, G. (2010). Identities, discourse and genres in corporate communication: An introduction. In P. Evangelisti Allori & G. Garzone (Eds.), *Discourse, Identities and Genres in Corporate Communication* (pp. 9–26). Bern, Switzerland: Peter Lang.

Fuoli, M. (2012). Assessing social responsibility: A quantitative analysis of appraisal in BP's and IKEA's social reports. *Discourse & Communication, 6*(1), 55–81.

Gabrielatos, C., McEnery, T., Diggle, P., & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, *17*(2), 151–175.

Ihlen, O. (2009). Business and climate change: The climate response of the World's 30 Largest Corporations. *Environmental Communication, 3*(2), 244–262.

International Corporate Identity Group, ICIG. (1995). *The 1st Strathclyde Statement on Corporate Identity*. Retrieved from http://www.icig.org.uk/the-strathclyde-statement (last accessed July 2017).

Jones, R. H. (2012). *Discourse Analysis. A Resource Book for Students*. Abingdon, UK: Routledge.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., Rychlý, p. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

Laine, M. (2005). Meanings of the term 'sustainable development' in Finnish corporate disclosures. *Accounting Forum, 29*(4), 395–413.

Laine, M (2010). Towards sustaining the status quo: Business talk of sustainability in Finnish corporate disclosures 1987–2005. *European Accounting Review, 19*(2), 247–274.

Lischinsky, A. (2010). The struggle over sustainability: A corpus approach to managerial conceptions of sustainable development. *Proceedings of the Critical Approaches to Discourse Analysis Across the Disciplines Conference*, 13–15 September, Łódź, Poland: Łódź University Press.

Lischinsky, A. (2011). The discursive construction of a responsible corporate self. In A. Egan Sjölander & J. Gunnarsson Payne (Eds.), *Tracking Discourses: Politics, Identity and Social Change* (pp. 257–285). Lund, Sweden: Nordic Academic Press.

Mahlberg, M. (2007). Lexical items in discourse: Identifying local textual functions of *sustainable development*. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora. Theory and Analysis.* (pp. 191–218). London, UK: Continuum.

Mair, C. (1997). Parallel corpora: A real-time approach to the study of language change in progress. In M. Ljung (Ed.), *Corpus-based Studies in English* (pp. 195-209). Amsterdam, Netherlands: Rodopi.

Mauranen, A. (2012). *Exploring ELF. Academic English Shaped by Non-native Speakers*. Cambridge, UK: Cambridge University Press.

Pilot, S. (2011). Companies are embracing corporate responsibility in their annual reports. Retrieved from http://www.theguardian.com/sustainable-business/blog/companies-embrace-corporate-responsibility-annual-reporting (last accessed October 2018).

Schlichting, I. (2013), Strategic framing of climate change by industry actors: A meta-analysis. *Environmental Communication, 7*(4), 493–511.

Selmi-Tolonen, T. (2011). Clean corporate citizenship identity. In V. K. Bathia, & P. Evangelisti Allori (Eds.), *Discourse and Identity in the Professions* (pp. 45–58). Bern, Switzerland: Peter Lang.

Swales, J. M., & Rogers, P. S. (1995). Discourse and the projection of corporate culture: The mission statement. *Discourse & Society*, *6*(2), 223–242.

Tadros, A. (1993). The pragmatics of text averral and attribution in academic texts. In M. Hoey (Ed.), *Data, Description, Discourse* (pp. 98–114). London, UK: HarperCollins.

Vodafone Group Plc. (2005). *Corporate Responsibility Report 2004/05*. Retrieved from https://www.vodafone.com/content/dam/vodafone/about/sustainability/reports/2004-05_vodafonecr.pdf (last accessed November 2017).

Helen Baker, Ian Gregory, Daniel Hartmann and Tony McEnery

Applying Geographical Information Systems to researching historical corpora

Seventeenth-century prostitution

Abstract: This chapter reports on research resulting from academics from linguistics, history and geography working together in order to cast light upon the geography of prostitution in seventeenth-century Britain. We will demonstrate the usefulness and untapped potential of combining corpus linguistics and Geographical Information Systems (GIS) as an approach to researching historical texts. Corpus linguists are beginning to pursue new methodological advances which encourage them to “think geographically” and provide opportunities to enrich their understanding of a body of texts by uncovering spatial patterns in types of discourse (Gregory & Hardie 2011: 298–299, 309). The ability to move from corpus text to a visual mapping of geographical data and then back into the corpus text provides rich opportunities for humanities scholars in general, and corpus linguists in particular.[1]

1 Introduction

The increasing digitisation of large bodies of historical texts offers researchers unprecedented opportunities to study, using very large volumes of material that were written in previous centuries. Large collections of transcribed historical

[1] The authors would like to acknowledge the generous support of the Newby Trust for the research reported in this paper. With reference to the work on GIS presented, the research leading to these results received funding from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant “Spatial Humanities: Texts, GIS, places” (agreement number 283850).

Helen Baker, Lancaster University, h.baker1@lancaster.ac.uk
Ian Gregory, Lancaster University, i.gregory@lancaster.ac.uk
Daniel Hartmann, Lancaster University, dhartmann2@gmail.com
Tony McEnery, Lancaster University, a.mcenery@lancaster.ac.uk

https://doi.org/10.1515/9783110489071-005

texts, for instance the Early English Books Online (EEBO) corpus[2] for the early modern period, can be searched by computer rapidly and accurately, revealing discourses of the past on a scale and depth never before practicable. These developments encourage collaborative research between scholars in different academic fields. Linguists, specifically corpus linguists, offer their skill and experience in manipulating and understanding large textual databases. Historians bring their specialist knowledge of the topic in question and its relevance in a wider historical setting. Historical geographers, meanwhile, can employ computing technologies to facilitate spatial interpretations of historical themes and events, giving historians the spur to review established historical theories and corpus linguists fresh methodological and analytical challenges.

The specific aim of this paper is to make full use of the 1.1 billion word Early English Books Online (EEBO) corpus to explore how much it is possible to discover about the differing experience and representation of prostitutes in different parts of Britain in the seventeenth century. This EEBO corpus contains the preponderance of sources printed in English between 1473 and 1700, although the majority of its content covers the seventeenth century. The version of the corpus we are using was created at Lancaster University based upon the work of the Text Creation Partnership. Our broader aim is both to show how we can use a range of techniques to account, in a large and complex corpus, for all of the mentions of a particular theme that are associated with places and to explore the discourses revealed, demonstrating how much they vary, if at all, between the different places.

2 Background to the study of seventeenth-century prostitution

In Britain the seventeenth century was a period of turmoil in every imaginable area – in religion, politics, economics and social structure – during which members of the lower classes paid a heavy price. Cities, particularly London, struggled to cope with the vast numbers of unskilled migrants who flocked to urban centres in the hope of finding work. Women found it almost impossible to acquire a long-

[2] See Early English Books Online http://eebo.chadwyck.com/home and the Text Creation Partnership http://www.textcreationpartnership.org/ This paper used version 3 of the EEBO corpus as created from the TCP data at Lancaster University. Subsequent versions of the corpus will see it further increase in size.

term position that paid sufficient to cover basic living costs. It is against this background that many women turned to selling the only commodity from which they could make a profit – their own bodies.

Prostitution is a highly relevant subject for present-day researchers and it illuminates many other areas of historical study. There were seventeenth-century sex workers from every social level – and probably from both genders – but the vast majority were women who entered the trade from poor agrarian or labouring backgrounds and did so as a result of severe financial hardship.[3]

Most prostitutes, therefore, were simultaneously members of two disempowered groups: women and poor people. However, throughout most of the seventeenth century, elite commentators assumed that prostitutes chose to sell their bodies because they were naturally lewd, promiscuous and immoral. It was not until the beginning of the eighteenth century that authority figures began to perceive sex workers with a sense of compassion and responsibility. By examining the experiences and textual portrayal of seventeenth-century prostitutes, we stand to gain deeper insight into the representation and culture surrounding both women and the lower classes in the period. Tracing elite attitudes to prostitutes, both in terms of the inconsistent legislative attempts to curb their business and their frequent appearances in broadsides, pamphlets and sermons of the period, has the potential to shed light on the development of moral principles in a wider cultural context and, in their expression, the workings of both national and local government.

Prostitution was considered to be an inappropriate and rather distasteful topic of study until the 1960s when a new interest in the lives of ordinary people, no doubt linked to the rise of feminist theory, emerged (see Attwood 2011, for information). Although there are very few secondary sources that concentrate on seventeenth-century prostitution, Mowry (2004) and Thompson (1979) have both examined pornographic literary works in the second half of the century, many of which reference commercial sex. Other monographs, although not covering our period directly, are also highly relevant, including Karras (1996) who has researched sex workers in medieval England, and Henderson (1999) who has analysed the legislative attempts to curb prostitutes working in eighteenth-century London. The vast majority of histories that do investigate seventeenth-century prostitution concentrate upon prostitutes working in London. Although much of this scholarship is accomplished, its findings cannot be automatically applied to the rest of the country. Indeed, many social historians are now turning to micro-orientated studies in order to present a more balanced and informed picture of

[3] See McEnery and Baker (2017) for a discussion of early modern male prostitution.

their period or area of specialisation. For instance, a number of scholars are choosing to examine the development and application of poor relief in various towns and parishes, many in the north of England, and this has contributed to a re-evaluation of the Poor Law.[4] Our understanding of seventeenth-century prostitution in England would undoubtedly be enriched by scholarship of a similar nature. Historians choose to study a specific location for a number of reasons. A certain area might fit in well with previous research interests or possess a weighty collection of archives waiting to be tapped. By making use of Historical GIS, we potentially have a stronger chance of identifying if there were other areas in the country where contemporaries believed that prostitution thrived and hence have a better idea where to direct future research. GIS mapping might, therefore, persuade scholars of the necessity of shifting the focus of their research away from the capital into unexplored but equally relevant places throughout the country. Contrariwise, it may also in part justify the focus on London. Importantly, by using such mapping we might begin to see whether discourses relating to prostitution are general, location specific or a mixture of both.

There is a dearth of records authored by lower class people in early modern England and this is overwhelmingly due to low literacy levels among the general population.[5] Even if someone living on the edge of poverty possessed literacy skills sufficiently high to allow them to read and write, many may have lacked the impetus or material resources necessary for them to record their lives. Any words of the poor were far less likely to have been valued, and therefore deemed worthy of preservation, after their deaths. Historians have, therefore, frequently turned to literary depictions of transactional sex, usually written by men of a higher socioeconomic class, in order to understand the experiences of prostitutes.

Seventeenth-century authors were fascinated by prostitutes. An array of broadsides, pamphlets, books, and periodicals, many reproduced in EEBO, referenced prostitutes and were perused by a new readership including tradesmen, merchants and skilled craftsmen. Rogue literature which purported to divulge the underground activities of real-life rogues, beggars and whores continued to be

[4] See, for example, Healey (2010), Hindle (2003), Stapleton (1993) and Williams (2011). The Old Poor Law (passed in 1597 and revised in 1601) introduced a rudimentary welfare system into England.

[5] Cressy (1977) has estimated that, in East Anglia, during the period between 1580 and 1640, 95 per cent of women were unable to sign their names. This decreased to 82 per cent between 1660 and 1700. He believes the proportion of adult female illiteracy in London between 1580 and 1640 stood at 97 per cent. However, this had dramatically decreased to 52 per cent by the 1690s while it persisted at around 80 per cent in rural areas.

popular. A number of prostitutes and bawds approached household-name status in the seventeenth century such as Mrs Cresswell, Elizabeth Holland and Damaris Page. References to real-life prostitutes also found their way into so-called 'ladies directories', for example, *A Catalogue of Jilts, Cracks & Prostitutes, Night-walkers, Whores, She-friends, Kind Women and others of the Linnen-lifting Tribe, who are to be seen every Night in the Cloysters in Smithfield, from the hours of Eight to Eleven, during the time of the Fair*, published in August 1691 which lists the names and attributes of twenty-two prostitutes.

The main challenge for historians hoping to make use of these literary sources is the issue of their credibility. Whose discourses are represented in such texts? Do these documents accurately reflect the lives and experiences of prostitutes of the period or did they simply pander to an audience which craved the thrill of reading about thieves, whores and bawds? Very little has survived concerning prostitutes before the eighteenth century other than court records which themselves introduce a new set of methodological challenges. Sharpe (1984) has written about the difficulties of building up a picture of crime in an area due to the array of different courts that were operating at the same time. For prosecutions of alleged prostitutes, the boundaries between jurisdictions were fluid which meant that some women were tried in secular courts and others were judged by church courts. There is a sizeable amount of indictment and recognizance records available which do contain some very interesting information such as where women accused of selling sex were based and if they were accused of another crime in addition to prostitution.[6] However, there are serious drawbacks to relying upon the records of court proceedings. One limitation is that our conclusions would be based purely upon prostitutes in relation to their alleged criminality. Many prostitutes were an accepted part of their community and did more with their lives than simply break the law. The voices of these women rarely emerge in transcripts and reports by court officials.

The main problem is one of coverage: most women who were detained on suspicion of trading as a prostitute were dealt with by informal legal procedures which often entailed little or no record keeping. People accused of committing a crime in early modern England were dealt with in various ways depending on the severity of the alleged offense. Felonies, such as murder or robbery, were heard at Crown courts which sat in the large towns. The Old Bailey, situated in London, was the most well-known of these courts. The vast majority of crimes were, however, regarded as misdemeanours and most perpetrators never reached court.

[6] See Walker (2003: 3).

Justices of the Peace instead preferred less expensive and quicker methods of dispensing justice such as mediation between victim and criminal. Sometimes an offender would be bound over by recognizance: this meant he or she would be ordered to appear at the next sessions to answer the charge but frequently the case did not proceed any further.[7] By far the most preferred method of dispensing justice to alleged prostitutes, and, for that matter, most people of limited financial means, was summary justice. In these cases, a pair of Justices, or sometimes an individual acting on his own, would issue a punishment, such as a spell in the houses of correction, a whipping or a fine, which would commence immediately. Robert Shoemaker (1991: 189) has shown that a woman accused of prostituting her body would most likely be sent to the houses of correction and that "loose, idle and disorderly conduct" accounted for around half of all commitments. The largest obstacle in using court records is that many women selling their bodies were never apprehended by an officer of the law so were entirely unrepresented in any kind of official documentation.

Another challenge to archival research is one of access and availability. The vast majority of court records remain undigitised and cannot be searched easily. The Old Bailey Corpus, a corpus derived from texts produced for the Old Bailey Online website,[8] does allow one to search through the proceedings of the criminal court, the Old Bailey. But its coverage starts three-quarters of the way into the seventeenth century. Another obstacle concerns the scale of the Old Bailey Corpus. It is a continuous sample throughout the period it encompasses, containing 2,163 volumes of proceedings covering nearly 200,000 words which has resulted in a corpus approaching 134 million words. However, relative to the span of time covered, the size of the corpus is fairly modest. While it is a very useful collection, the Old Bailey Corpus is probably more suited for a study of frequent features, such as grammatical features or concepts frequently addressed in the courtroom.[9] We have also seen that while felonies and the more serious misdemeanours, such as rape, were tried at the Old Bailey, most cases involving prostitution never reached the court.

So when we consider the EEBO corpus, a provisional answer to the question of whose discourse is present in the corpus is 'almost certainly not that of prosti-

[7] Nightwalkers and bawds were sometimes prosecuted by indictment because they were perceived as disturbers of the peace.

[8] Old Bailey Online - The Proceedings of the Old Bailey, 1674 to 1913 http://www.oldbaileyonline.org/

[9] See for example Traugott (2010).

tutes themselves'. The texts are likely to be about prostitutes, where they are mentioned, not by prostitutes. Hence any discourses revealed are likely to be those of 'others', not prostitutes themselves. But to explore this hypothesis we need first to explore texts mentioning prostitution in the EEBO corpus.

3 Choice of search-terms

ape-gentlewoman, commodities†, courtesan*, Covent Garden ladies†, crack, daughters of sodom, doves of Venus, girls of pleasure†, harlot*†, hydra, jade, jilt*†, kind woman, lewd / wicked / abominable / debauched / disorderly person* or woman, miss*†, monster*, nightwalker*†, nocturnal privateers, nymph*†, pirates of the night*, prostitute*, punk*†, quean*†, she-friend*, strumpet*, the sisterhood, traffic*†, trull*†, wanton ladies, whore*†, woman of iniquity, woman of the town*

Fig. 1: Possible terms to describe prostitutes in the seventeenth century. Words or phrases occurring in the OED are marked with an *, words or phrases occurring in the *Chambers Slang Dictionary* are marked with a †. Words or terms the historian thought were rare when describing prostitutes are underlined.

Accessing what is being said about prostitution in the seventeenth century is not necessarily easy. Discussions about mainly fictional prostitutes and warnings to avoid their company are present in EEBO sources, but these are fragmented due to the very large volume of material that EEBO contains. We therefore have to decide which search-terms to use to identify relevant passages in the texts. While using intuition is tempting, this would be to overlook the fact that there are no living speakers of early modern English whose language we can observe and judge our intuitions against. Some of the nouns we might suggest for *prostitute* may not have been in use four hundred years ago or, indeed, even twenty years ago. Dictionaries such as the *Oxford English Dictionary* (OED), the *Chambers Slang Dictionary*[10] (Green 2008) and the *Historical Thesaurus of the Oxford English Dictionary* do include words relating to seventeenth-century prostitutes but to use these sources efficiently we must already know what terms we want to check and rely on the date feature of a dictionary like the OED. In its online version, the OED

[10] Green (2011) has also published a specialist volume on the language of crime which devotes a section to the various nicknames prostitutes have received throughout the centuries.

does allow researchers to review words occurring in definitions within a particular period so we can develop, for instance, a list of words and phrases which originated in the sixteenth century to act as a guide by looking for *prostitute* in the definition. However, the OED is a weaker source of slang words than a slang dictionary so it would be necessary to use both sources concurrently and neither would indicate which words are worth examining in detail. Another source of lexical insight relates to the intuitions of native speakers of the seventeenth century. *Lexicons of Early Modern English* (LEME, Lancashire 2015) is a digital database of a wide variety of historical dictionaries that researchers are able to search, including head words as well as definitions and limiting the date if desired. However, we must approach these dictionaries with caution as they pre-date important developments in lexicography of the eighteenth century that led to data being gathered more systematically and presented in a standardised format to a greater degree. Yet we need not set aside such lexicographic resources – we can use the historian's intuition and corpus evidence to look at the dictionaries to get a sense of which words that the historian knows or that the corpus attests may be new to the dictionaries. Hence Figure 1 shows a range of possible words or phrases associated with prostitution along with their source.

In Figure 1 the historian knew some phrases not known to the dictionary – for example the rare *doves of venus*. Indeed some common forms, attested in the corpus and known to the historian, were absent from the dictionaries consulted, notably *jade* which occurs 1809 times in the EEBO corpus in the seventeenth century. After a study of the words used to refer to prostitutes in the seventeenth century, we focussed upon a number of frequent words to look at in our GIS study. We need frequent words because the mapping algorithm links examples of a word to locations mentioned in the context of that word. As not every mention of a word is linked to a location, we needed to focus on frequent words so that a reasonable subset of examples could be discovered where that word was linked to a location. We will return to this point shortly to illustrate it. Another option would be to compile a list of words from a specific text of the period which is known to be a rich source for the topic of prostitution. The newsbook *Mercurius Fumigosus*, produced from 1654 to 1655 by the Royalist writer John Crouch, made regular references to prostitutes in a very informal manner and is a good example of such a text. However, again this is problematic because we may only end up with words that were mostly used in the 1650s or only used by Crouch himself.

A more useful approach when embarking upon a corpus linguistics analysis of a historical subject is to utilise the traditional skills of the historian. Being familiar, by virtue of close reading of an array of early modern texts, historians are able to suggest what words and phrases were used to refer to prostitutes in the period in question and whether these were commonly used or not. For the purpose of this study, a long list of words was drawn up by an historian and then passed to a corpus linguist who, after analysing the words against the EEBO corpus, made the final decision of which would be most productive for GIS mapping.[11] The words collected by the historian included a number that are probably familiar to present-day readers such as *courtesan*; *harlot*; *nightwalker*; *prostitute*; *strumpet*; and *whore*. Other words were far less familiar: *ape-gentlewoman*; *commodity*; *crack*; *jilt*; *monster*; *quean*; *she-friend*; *trull* and *wanton lady*.[12]

The corpus linguist is faced with a number of issues when undertaking an analysis of these words. We decided not to pursue the very rare words, i.e. single instances which refer to prostitutes because, although interesting, they can at best provide us with an expression used by one writer. Of more interest were words such as *strumpet*, which are moderately frequent[13] and hence can be explored relatively easily. The words that appeared more frequently in our corpus present a greater degree of challenge. *Monster*, for example, occurs 12,338 times in the seventeenth-century material of the EEBO corpus yet the vast majority of these matches do not refer to prostitutes. Checking each of these concordances one-by-one to ascertain whether they referred to the practice of prostitution would be a painstaking and lengthy procedure. At this point, corpus linguistics can help again – collocation allows the analyst to make a reasoned decision about whether or not to pursue frequently-occurring words whose context was unclear. In the case of *monster*, collocation revealed that *woman* was the only collocate relevant to the study of prostitution. In a handful of cases this did indeed lead to a concordance referring to prostitutes but this was extremely rare hence the word *monster*, while referring sometimes to prostitutes, could be set aside.

Using the tools of corpus linguistics allowed us to focus our study on four words which are most often used to mean prostitute and which, moreover, occur

[11] This approach has much in common with that of Nevala and Hintikka (2009) who have examined thirteen pamphlets available in EEBO and the Eighteenth Century Collection Online (ECCO) in order to trace changes in terminology used to refer to prostitutes throughout the seventeenth and eighteenth centuries. Taavitsainen (2016) sorts terms of address into different semantic fields and offers evidence of a metaphorical connection between a number of feminine terms and prostitution. *Mistress*, *miss* and *madam* are sometimes used pejoratively to suggest the addressee has links with prostitution.

[12] For a more detailed discussion of how we compiled and shortlisted words and phrases used to refer to prostitutes, see McEnery and Baker (2016).

[13] There are 2,744 examples of the word *strumpet* in the seventeenth-century material in v3 of the EEBO corpus.

frequently in the corpus. These words were PROSTITUTE, which occurs 5,374 times, HARLOT 14,995 times, STRUMPET 4,444 times and WHORE 33,678 times.

4 From search-terms to maps

The next challenge is to add geographical information to this such that we can identify and map the places that are being associated with prostitution. This involves using a process called geo-parsing.[14] This is a two-stage process. In the first stage natural language processing (NLP) techniques are used to identify candidate place-names. In the second stage co-ordinates are found for these by matching the candidates to a gazetteer, which is effectively a database table that lists place-names and their co-ordinates. Geo-parsers, such as the Edinburgh Geo-Parser (Grover et al. 2010), are available that geo-parse an entire corpus. In this case, however, geo-parsing the entire corpus was not felt to be appropriate for reasons associated with both accuracy and processing times. From an accuracy perspective, even when relatively modern corpora are geo-parsed the results will be error-prone (Tobin et al. 2010). For earlier corpora, where reliably identifying place-names can be expected to be more complex, errors are likely to be more significant. Evaluating how many errors there are, correcting them where possible, and evaluating the impact of the remaining errors in a corpus of over a billion words is likely to be problematic. From a processing perspective, the time taken to geo-parse a corpus of this size is also likely to be prohibitive. Instead we use a process called ‘concordance geo-parsing’ (Rupp et al. 2014). This starts by using corpus linguistics software to extract each occurrence of a search-term and its ‘concordance’, the text that surrounds it. In this case we used a wide concordance of fifty words to the left and right of the search-term to give the geo-parser plenty of contextual information. This text was then geo-parsed in the usual way using the Edinburgh Geo-parser. The results were then explored for errors. Importantly, the corrections to these in the form of additions, deletions or changes, were manually written to an updates file. If the geo-parsing is repeated, for example with a new search-term, these corrections can be automatically included in the results. By starting with a relatively rare search-term and working up to more common ones this allows relevant material to be geo-parsed, checked and corrected such that we can be confident of their accuracy.

[14] While the system used in this paper is not publicly available, for readers interested in experimenting with geo-parsing, systems such as http://cliff.mediameter.org/ are freely available.

Concordance geo-parsing bears some similarities to, yet is different from, collocation in corpus linguistics. For example, while both use spans of words for the purposes of analysis, in corpus linguistics collocation uses a relatively short span of words, typically five words left and right, to look for quite localised relationships. Yet the span usually used to explore collocation seems too restrictive for geo-parsing, as noted. Secondly, geo-parsing is simply looking for word co-occurrence in a given window. It wants to find all examples of a place-name, for example, co-occurring with a word that the user is interested in. Questions of statistical significance and effect measures are, rightly, set aside. To exemplify both points, if we use a five word left/right span for a word like WHORE for example, then, using a log-ratio statistic, we find only one collocate which appears to link the word to a location *israel*, which collocates with WHORE 112 times. As this location is not based in the UK, this approach would have yielded this study no useful data. While the collocation statistic allows us to search the many examples of WHORE quickly, it encourages us to look for place-names that associate frequently in close proximity to WHORE – this is not quite what we want. We want any place-name associated with WHORE. Using the geo-parsing and looking for locations in Britain by applying a wider collocation window and no measure of collocation strength/significance yields 269 good examples of WHORE where the word is clearly linked to a place-name in Britain. Given that our goal is to find the locations that the words we are interested in are linked to, our method in this chapter is to use geo-parsing, a wider collocation window and an approach to co-occurrence in which strength of association is discarded as this clearly produces many more useful examples than would normally be produced through a standard approach to collocation.

The aim of this process is to provide a geographical location for every mention of the words we are searching for in the corpus. What it actually does is to provide a co-ordinate-based location for every place-name identified by the geo-parser that lies within fifty words left or right of one of our four search-terms. As we have co-ordinates, this information can be read into GIS software for mapping and subsequent analysis. Initial mapping reveals that further correction is required primarily because terms for prostitution are often used as insults and these can and do collocate with place-names. This leads to major clusters in places such as Babylon and Rome which are driven by "whore of Babylon" and Reformation-era insults aimed at the Catholic Church which is frequently described as the *whore of Rome*. These two phrases are even sometimes run together such as "...did both know and confess the Church of Rome to be the Whore of Babylon" (Floyd 1612). Thus the mapping reveals something that is interesting from both a linguistic and historical perspective, but which tells us nothing about the geography of prostitution. Exploring the maps further revealed that very few mentions of places outside Britain accompanied genuine references to prostitution, hence making our focus on Britain more strongly warranted.

The corrected distribution of the four search-terms combined is shown in Figure 2. The map uses two different ways of representing places that co-occur with the search-terms. As well as using point symbols, a density smoothed surface is also used such that darker shading represents places where more points occur close together. This approach makes the overall pattern clearer than simple point mapping and is particularly effective at drawing attention to spatial clusters. What seems apparent from the figure is that there is a clear concentration of these words being used alongside place-names of urban centres. London stands out, while in the north of England it is easy to identify hotspots in Manchester, Leeds and York. One problem with Figure 2 is that it might be speculated that the pattern it shows reflects the distribution of place-names in the corpus as a whole rather than the distribution of the search-terms. If most of the books from the corpus were published in London and other urban centres it would follow that these places are likely to be the most mentioned in the texts and thus the pattern shown in Figure 2 may be a random artefact of this rather than telling us anything about prostitution. One way to compensate for this would be to compare the spatial pattern of prostitution instances to the underlying geography of place-names from the corpus as a whole. Statistical tests such as Kulldorf's Spatial Scan Statistic (Kulldorf 1997) are available that allow us to identify clusters of places where there are more or fewer points when compared to a background geography, termed 'hotspots' and 'coldspots' respectively.[15] This approach has been applied to identify clusters in texts (Murrieta-Flores et al. 2015; Gregory & Donaldson 2016). However, in this case it cannot be easily applied as the background geography of the corpus as a whole is unknown because of the problems of geo-parsing the entire corpus described earlier.

[15] In this analysis this was implemented using SatScan, see: http://www.satscan.org (viewed 24th Oct 2014).

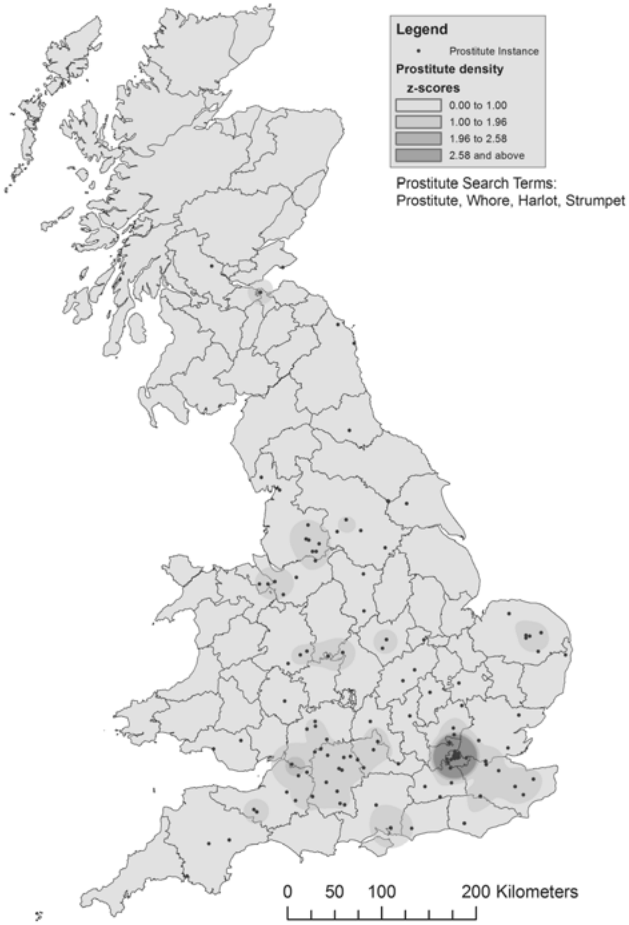


Fig. 2: Places that co-occur with *whore*, *harlot*, *strumpet* and *prostitute*.

Rather than test against the entire corpus, a modified approach was used in which two themes, women and urban, were decided on whose geographies could be compared to the prostitution pattern. While these two themes are somewhat arbitrary, they do allow us to compare the geography of prostitution with a measure of the urban geography and a measure of the geography associated with women to see where prostitutes were referred to more or less often than would be expected. For urban areas the terms *market, church* and *house* were selected and geo-parsed using the concordance geo-parsing approach described above, for women *mother, widow* and *nun* were used.

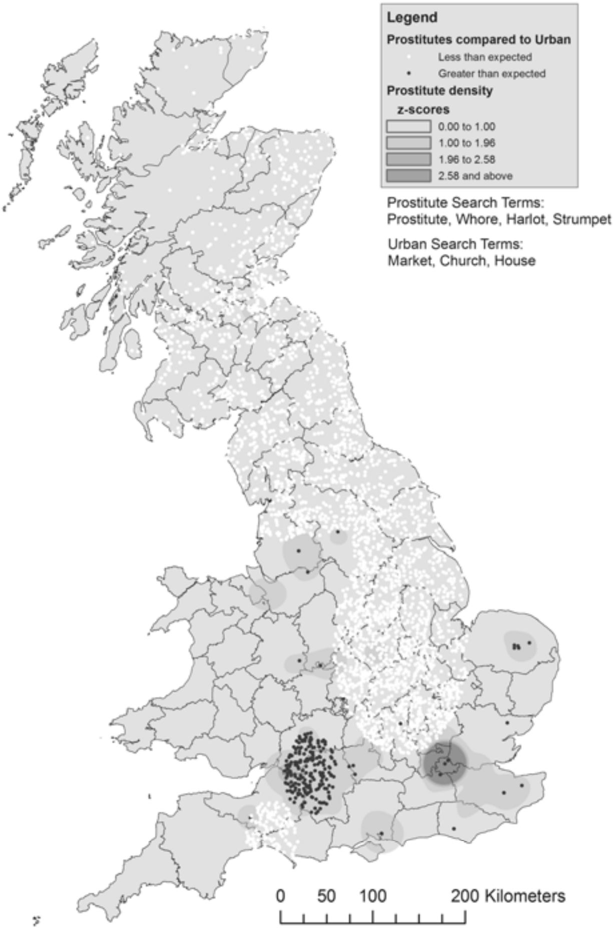


Fig. 3: Hotspots and coldspots of prostitution instances when compared to instances related to urban areas. Hotspots are where observed>expected, coldspots are where observed < expected.

Figures 3 and 4 show the results of using Kulldorf's Spatial Scan Statistic to compare the geography of prostitution with the patterns for search-terms associated with, respectively, urban areas and women. Hotspots, places where there are significantly more prostitution instances than the background pattern would have us expect, are shown as black points over the density smoothed pattern taken from Figure 2, coldspots are shown as white dots. The patterns shown in the two maps are broadly similar with London, parts of the West Country, and various smaller urban centres emerging as hotspots on both. Much of the rest of the country appears to be coldspots revealing that references to prostitution are concentrated in the south and west of England. The fact that the patterns in Figures 3 and 4 are very similar despite being based on very different sets of background search-terms suggests that the areas identified as hotspots are in fact the areas that are closely associated with references to prostitution once the variable geography of the background corpus is taken into account.

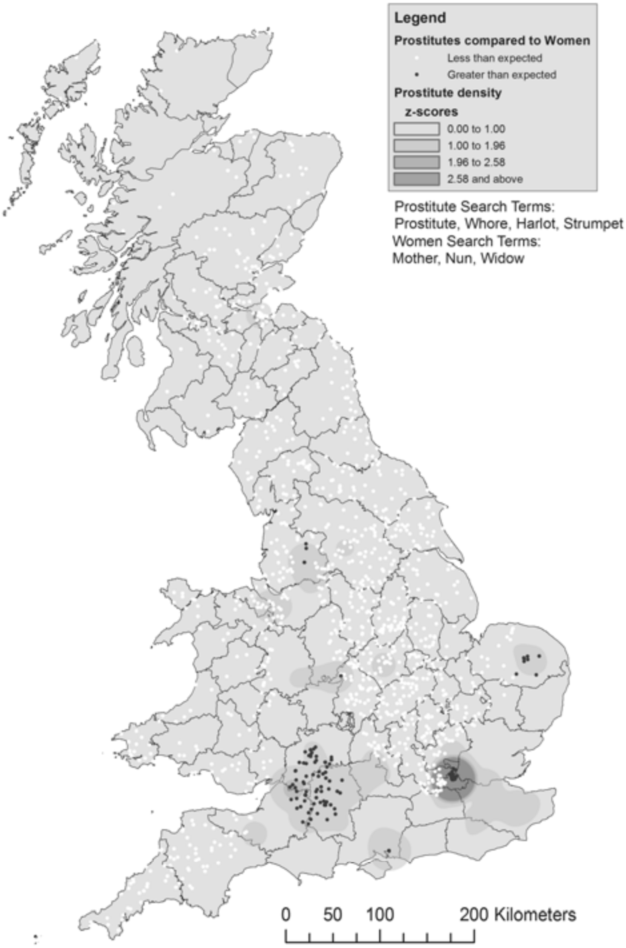


Fig. 4: Hotspots and coldspots of prostitution instances when compared to instances related to women. Hotspots are where observed > expected, coldspots are where observed < expected.

5 Interpreting the patterns

What do we have to gain from these maps? The most obvious benefit is that they allow us to visualise large collections of data rapidly and easily. Historians traditionally make decisions about the direction of their research or its overall arguments by closely reading relatively small proportions of large collections of text. The rise in digitisation of historical texts means that there is far more information available to scholars than can be easily read – historians are, after all, limited by how much they can read and absorb in one working lifetime. GIS mapping is one of the tools now available which allows scholars to reap the benefit of having access to large amounts of raw data without being swamped by its content. It effectively summarises and presents information so that scholars can, among other things, make a more informed choice about where to focus their reading. The same is also true of corpus linguists – the maps help guide their analyses of the words in question for example, setting the challenge to understand the role of urban populations in discourse. In addition, however, there is the possibility that such maps may allow us to more accurately identify geographically bound discourses. As noted, geo-parsing provides more data linking words and locations than collocation appears to. This opens up the possibility that we may see more clearly the intersection of space and discourse. We will explore this hypothesis shortly.

Let us now look more closely at the texts that created the patterns shown in the maps. We can see that there are a number of intriguing hotspots where prostitutes are mentioned more than we would expect in comparison to our standard "urban" or "women" words. These include London, Manchester, Norwich, Birmingham, Colchester, Portsmouth and a large cluster in the West Country. The London cluster is perhaps predictable based on the historiography, but one of the interesting things about maps such as these is that it reveals other areas that may be worthy of further investigation.

The large cluster in the West Country seems to be associated with anti-Catholic sentiment in which allegations of sexual impropriety are made against high-ranking churchmen. For instance, Francis Fullwood's *Church-History of Britain* contains a list of unchaste abbots and their place of residence and invites his readers to read, blush and sigh. He tells us that in Bath Monastery, Richard Lincombe had seven whores and was also a sodomite, while in Abingdon Monastery, Thomas, the Abbot of Abington, kept three whores and had two children by his own sister (Fullwood 1655). Other texts contain similar lists: "[a]t Bath, one Monk had seven whores…" (Care 1679); and "[i]n the abbey of Monkenferlege in Salisbury diocese, Levvis the Prior had 9. harlots, Richard the Prior of Maiden Bradley had v. harlots and six bastards” (Bale 1574). These texts reveal how, in a period of intense anti-Catholicism, unmarried women engaging in sex were identified as whores and were used to slur the reputations of high-ranking ecclesiasts. References to *whores* and *harlots* illustrate the ambiguity of the terms – it is more likely that these women, if they existed, were labelled whores because they were believed to have consented to non-marital sex rather than because they accepted money in exchange for sex. While this is interesting in terms of shedding light upon how extra-marital relationships were imagined by contemporaries and, particularly, how women were condemned for their involvement in such liaisons, it does not reveal a great deal about the practice of early modern prostitution.

The smaller concentration of matches in the Portsmouth area can also be readily explained. A collocate of Portsmouth in the EEBO Corpus is *whore*. Yet in this case concordance analysis reveals that *portsmouth* specifically refers to Louise de Kérouaille who was granted the title Duchess of Portsmouth by her lover, Charles II. French and Catholic, Louise was the subject of immense popular contempt. The publication of a tract by her enemies at the beginning of 1680 entitled *Articles of High Treason and Other High Crimes and Misdemeanours against the Duchess of Portsmouth* accused Louise of twenty-two acts against the Kingdom, including being a participant in the Popish Plot, a fictitious conspiracy imagined by Titus Oates which intensified anti-Catholic hysteria (see Wilson 2004; Conway 2010: Chapter one). Oates (1696: 70) later wrote in a condemnatory biography of James II:

> And this remember, Sweet Sir, that the Whore Portsmouth, with her Bastard Son, was by your Royal Brother sent to France, to renew the Dover Treaty, in which she was more successful than your Sister of blessed memory, for she was caressed by the French King, and her Bastard honoured as a Prince of the Blood, and she settled a firm Correspondence between Lewis of France and Charles of Great Britain, pursuant to the Treaty at Dover.

This swipe at Portsmouth supports a recent historical re-evaluation of the Duchess as possessing considerable political agency, controlling advisors’ access to the King and negotiating the rise of key figures. Again, although the maps have not led us to prostitution as it was practised in the back alleys of cities and towns, the charges against de Kérouaille are of considerable relevance in terms of giving us a flavour of the major political and religious issues dominating post-Restoration England. On a slightly different note, this also raises the issue of the ambiguity between people’s names and place-names in this period. It could be argued that in this case *Portsmouth* is entirely associated with a person and not the place. However, in the cases given above, people such as Thomas, Abbot of Abingdon are clearly associated with a place, making this a somewhat moot point.

Other clusters outside London reveal less. Norwich reveals little of any substance, Colchester is largely the result of a story about a whore who was hanged for murdering her children (Bunyan 1680), and in Canterbury we return again to reports of monks and their relationships with prostitutes.

So the maps seem to align with the patterns of historical research – the largest main hotpot in Figures 3 and 4 is the area around London. In and around the city, prostitutes were being mentioned more than we might expect in comparison with our "urban" and "women" search-terms. Exploring the corpus reveals a variety of discourses that create this emphasis. The first is writers complaining about the prevalence of prostitution in the city, for example, in excerpts from two sermons by Hugh Latimer: "…how GOD is dyshonored by whoredom in this city of London" (Latimer 1549) and "there is now more whore doom in London, then ever there was on the bank" (Latimer 1562). The second, perhaps developing this further, draws explicit links between prostitution and crime either in generalities – "conceive what Outrages and Uproars would be in London, with Whoring, Thieving, Plundering, if there were no Government to restrain" (Lightfoot 1700) and "[s]easoned with the fees and bribes of all the whores and thieves that live in Westminster, Coven-Garden, Holborn…" (Milton 1642) – or specific cases "Thomas Savage… in the Parish of Stepacy, near London, who by the instigation of a Whore, blinded with lust, was wrought upon to Murder his fellow-servant" (Anon. 1668). The third discourse is concerned with men's relationships with prostitutes and their wives such as "…do fear least your wife should know that you keep a whore at Putney" (Puritanomastix 1642), and "[m]y Husband's leawd, given to go astray … There's not a Whore in London, nor about, But he hath all the haunts to find her out" (Rowlands 1613). These two quotes reflect the fear and shame frequently associated with relationships with prostitutes and the linking of that relationship to the wife of the client.

Many of these discourses unsurprisingly emerge from religious works. People living in early modern England were preoccupied with religious issues and, as a result, a large proportion of the texts within Early English Books Online are religious in essence. However, our search-terms appear fairly evenly in a variety of other texts, including true-life crime accounts, satires, dictionaries, histories and transcripts of defamation court cases. After the Restoration of 1660, there is an increase in plays, specifically, comedies, which have characters who are inherently linked to whores or whoredom – for instance, Sir Oliver Whorehound in *The life of Mother Shipton* (Thomson 1670) – or which accuse both male and female characters of sexual transgression: "[c]ould not you have lived better at home, by turning thy Whore into a Wife, than hear by turning other Men's Wives into Whores?" (Farquhar 1699) and "[t]o Bridewell you prostitute" (Ruggle 1662). The playhouse underwent a change of fortune with the return of the monarchy and a stream of comedies which made fun of sexual misbehaviour and hypocrisy proved very popular among playgoers. At the moment, in order to identify patterns of social attitudes within certain types of printed works, we must rely on manual sorting. However, work is currently being undertaken to sort the texts contained within the EEBO corpus into literary genres which would enable this process to be achieved with greater accuracy and rapidity.

The mapping technique is helpful in revealing these discourses, but let us now return to a question posed earlier: to what extent are the discourses geographically bounded, as the analysis suggests? To test this we looked at collocates[16] of the word which contributes the bulk of examples, WHORE. In the case of the first discourse, there are no collocates of the word that imply a problem with the number of whores in general in Britain, adding weight to the argument that what the map reveals is a location specific discourse – the perceived problem of there being too many whores was confined to London. By contrast, WHORE has a host of collocates linking the word to crime. In the top 200 collocates for WHORE collocates link it to cheating (*jilting*), theft (*thieving, theft, stealing, thievery, thievish, pilfering*) and illegal sexual acts in addition to prostitution (*buggery*, *incest*). So this discourse, while present in London, seems much more general in the language. The same is true for the third discourse: collocates of WHORE clearly show that the wife of the client and the prostitute are often framed together in discourse – in the top 200 collocates, words such as *adultery*, *spouses*, *adulterer* and *adulterous* demonstrate this clearly. So while useful in suggesting localised discourses, careful checking back with the whole corpus is necessary to differentiate discourses that appear localised, but which are in fact general, from those which truly appear to be localised in nature. However, there is no doubt in this case that geo-parsing was useful in identifying a localised discourse that would not have been revealed easily using tools available in standard concordance packages.

How justified were the writers in claiming that there was an issue with the numbers of prostitutes in London? The existence of the London brothels, particularly in Southwark, certainly increased the visibility of sex workers and the social issues surrounding them. There have been various estimates of the numbers of prostitutes that were operating in seventeenth-century London. *The Ladies Champion* of 1660, for instance, gives a figure of 1,500. In the same year, *The Practical Part of Love* asserted that a list of whores' names would cover thirty pages.

[16] To explore these we used the log-ratio statistic, using a collocation span of five left and right.

Thompson (1979: 57) has calculated this would make a total of 3,600 from a London population of around 350,000.[17] When London is compared to other cities of the time in the British Isles, this localised discourse was arguably justified because the nature and scope of prostitution in London did differ from the trade of sex in other places.[18] There were certainly more brothels in the capital than anywhere else in the country. Many were highly-organised establishments where a client could buy time with a woman who had received instruction concerning personal hygiene, social grace, and sexual prowess. Prostitutes working for the brothel-owner Charlotte Hayes, for instance, were given music and dancing instruction and were taught to carry themselves in a ladylike manner. At the other end of the market, madams employed different tactics in order to extract as much money as possible from clients. Inhabitants of lower end establishments would apply a heavy layer of make-up and encourage their customers to buy alcohol as well as sex. Linnane (2003) has estimated that in some brothels, the women would be expected to service over fifty men each night.

The majority of prostitutes working in the capital did not work in brothels. Many rented a room in a lodging house which was crammed full of other working women while others hired a room in an "accommodation" house where they could entertain clients away from the eyes of constables and beadles. At the bottom of the whore hierarchy were the street-walkers who conducted their business in back alleys, dark streets, one of London's parks or any convenient location. Some of these women were homeless and slept in barns or sheds when not working. In March 1762 James Boswell wrote in his diary of engaging with one such prostitute during a stroll in the park, "whom without many words [I] copulated with free from danger, being safely sheathed. She ugly and lean and her breath smelled of spirits. I never asked her name. When it was done, she slunk off" (quoted in Ackroyd 2000: 374). Exploring the corpus to find buildings and places with which prostitutes are linked supports this view – we looked at all four words investigated through the map and looked at the top 200 collocates of those words in the EEBO corpus. While *brothel* and *brothelhouse* are collocates, so are *stews* and *play-house* indicating a wider association in the first place between prostitu-

[17] For further estimates of the scale of prostitution in the period see McMullan (1984) and Denlinger (2002).

[18] Griffiths (1998) has investigated the notion of a localised discourse in his research on the changing meaning of the term *nightwalker*. Using archival sources from the London Bridewell, a prison-hospital, he shows that the word was increasingly feminised in the seventeenth century and argues that this reflected a desire to curb female sexual transgression in a climate of fear and recidivism.

tion and poverty and in the second between prostitutes and playhouses, a location claimed by commentators of the time to be linked to the practice[19] and acknowledged by historians as a place in which solicitation and the sex trade took place in the era (see Ditmore 2006: 31, for a brief account).

The second discourse identified in the London texts concerned the relationship between prostitution and crime, which, as noted, is well evidenced through collocation. Yet the link between the two has been explored by present-day historians who have argued over the extent to which a criminal underworld existed in urban areas in the seventeenth century. John McMullan, for instance, has written about a gangland culture in which pick-pockets, cutpurses, thieves, fences, panderers and prostitutes took advantage of the space and anonymity that London offered (McMullan 1984: 1–3, 15). Brothels did provide meeting-places for criminals who wanted to plot their next crime and some bawds offered a brokering service whereby they helped dispose of stolen goods. Court records prove that some prostitutes regularly picked pockets and that others worked alongside male criminals in order to blackmail or extort money from clients. However, it appears that such conspiracies were informal and opportunistic and they often caused more trouble for the prostitute than benefit. Given the dynamics of two liminal cultures co-existing on the margins of society and the practices outlined above, it seems hardly surprising that the link between prostitution and other forms of wrongdoing is a general, rather than London specific, discourse. Indeed the Society for the Reformation of Manners, operating from the last decade of the seventeenth century, believed that bawdy houses in general, not simply London bawdy houses, were “not only the nurseries of the most horrid vices, and sinks of the most filthy debaucheries, but also (as we suppose them) the common receptacles, or rather, dens of notorious thieves, robbers, traitors and other criminals” (*Antimoixeia* 1691, quoted in Shoemaker 1991: 249).

With reference to the third discourse, literature from the early modern period offers many examples of wily and malevolent prostitutes who led decent husbands and male servants astray. Griffiths (1993) has explained how prostitutes were despised because they represented a threat to the stability of the established social order. They were thought to have turned their back on religion and rejected family life, despite many having children of their own. Likewise, a husband who spent time with a prostitute was perceived to be neglecting his family and betraying his wife (Griffiths 1993: 40–41). These discourses were very much prevalent in the public sphere. There are many instances of communities working together

[19] The claim that playhouses and prostitution were linked was popularised by the work of Stubbes (1583).

in an attempt to remove prostitutes from neighbourhoods. During the trial of Elizabeth Holland, people living near her brothel at St Leonard's in Shoreditch informed the court that they were being disturbed well into the night by clients partying, swearing and drinking. A specific complaint was that young girls living in the area, daughters and household servants, were frequently being mistaken for Holland's whores (Burford & Wotton 1995: 80).

6 The London focus

Given the results of our mapping, it is understandable why most histories concerning early modern prostitution have concentrated on London. Indeed, their authors were following a trail left by literary sources of the time: Arnold (2011), Linnane (2003) and Emerson (2002) have all published books on the practice of commercial sex in the capital that are well-informed by contemporary texts. The information provided by the mapping has shown definitively that writers living in early modern times chose not to write in detail about the experiences and portrayal of prostitutes residing outside of the capital – all they have left us are some ambiguous references to the existence of whores in a scattering of towns and cities. This is all the more frustrating because we know that prostitution did exist throughout England and Scotland and occasionally we catch a glimpse of it in areas outside of London by means of other sources. The magistrate, Robert Doughty, for instance, investigating pauper apprenticeship in Norfolk in the 1660s concluded that the most able girls abandoned their apprenticeships because they could earn more "by spinning and knitting, gleaning & stealing in harvest, & perhaps by secret whoredoms all the yeare" (Hindle 2003: 222). Norwich, the country's second largest city by the seventeenth century, is suggested as a potential hotspot by the maps but a qualitative analysis of the matches revealed nothing of interest. Norwich has long been of interest to historians: inhabitants of the city were heavily involved in Kett's rebellion[20] and, prior to the Poor Law introduction of 1597 and 1601, Norwich was the first provincial city to introduce compulsory rates which funded a rudimentary poor relief system. However, no focused research concerning early modern commercial sex in Norwich has been undertaken. We do know that bawdy houses existed in the city because the Corporation unsuccessfully tried to close them down in 1668 and 1681 (Corfield 1972).

[20] A revolt in 1549 which began in Norfolk in which attempts by wealthy landowners to enclose their land with fences was resisted.

Humphrey Prideaux visited Norwich in 1681 and declared: “[t]his town swarms with alehouses, and every one of them they tell is alsoe a bawdy house”.[21] The Norwich census of the poor of 1570 named eleven harlots living in the city although it appears that some of these women were simply unmarried mothers (Pound 1971: 76, 82, 92, 93).[22]

Very little research has been conducted into prostitution in towns but there are occasional descriptions of women selling sex in such places: eighty such women were identified in Leith in 1692 and the town of Deal in Kent was described as being residence to twenty-six whores in 1703 (Thompson 1979: 59). Rural prostitution has also been neglected by scholars but to a lesser degree. G.R. Quaife’s *Wanton Wenches and Wayward Wives: Peasants and Illicit Sex in Early Seventeenth Century England*, which is based upon court records from Somerset, discusses pre- and extra-marital sex away from the capital (Quaife 1979).[23] Quaife describes how in small rural areas some whores were accepted by their neighbours and tolerated by village officials. One type of prostitute, for instance, operated discreetly from her own home and sometimes accepted food or labour in exchange for sex. Some of these women were married and engaged in a little casual prostitution with or without the consent of their husbands. Sometimes an alehouse barmaid would offer sexual favours on the side if the opportunity arose. Wandering female vagrants often passed through towns and villages selling sex: the village authorities were more concerned with preventing such travellers settling in the area and claiming poor relief, so they did all they could to encourage them to leave rather than detain them (Quaife 1979: chapter six). Wrightson (1973) who has undertaken regional histories of early modern Essex and Southwest Lancashire, concludes that the few mentions of prostitution in court records of the areas stand out because they are so rare. Perhaps, then, by failing to highlight occurrences of rural prostitution, the maps do reflect the rarity of professional prostitutes in these areas.

In summary, on the basis of the literary and corpus evidence, the focus on London in studies of seventeenth-century prostitution seems justified. While prostitution occurred elsewhere and, perhaps, was more tolerated because it was

[21] Thompson (1979: 55–60) writes that alehouses in large market towns, busy seaports and important communications junctions were perceived to double as brothels.

[22] One woman, Jone Skyner, is described as a *grese mayde* or grass maid, a mother of a child born out of wedlock within a household entry, but at the end of the ward section she is noted to be a harlot. She may have simply been termed a prostitute for being known to have engaged in non-marital sex. Another woman, Mable a Breten, is similarly noted to be a whore but is earlier described as a servant. Also see Silvester (2012).

[23] Shoemaker (1991) covers an area abutting London as well as the capital itself.

less salient, there does seem to be evidence that prostitution in London was unusually widespread relative to the other areas in the British Isles.

7 Conclusions

This chapter has shown how GIS and corpus linguistics can be used to explore a very large corpus in relation to a theme that is referenced relatively rarely and where the references that do exist are fragmented throughout the corpus. The study shows that London is the place that is most closely associated with prostitution and within the city there are three main discourses: those associated with complaints about there being too many prostitutes in the city, those that link prostitution to crime, and those that talk about people's relationships with prostitutes with fear either of a man's relationship with a prostitute being discovered or of a woman finding her husband being involved with prostitutes being particularly common. Outside London the most common discourse related to prostitution is largely driven by anti-Catholic sentiment particularly aimed at monasteries, abbots and monks. Surprisingly little of substance emerges about prostitution in other parts of Britain and almost nothing in rural areas.

Let us conclude by considering two questions. Firstly, let us return to the question of whose discourse we are exploring. Secondly, what does the inter-disciplinary focus of this work yield? To address the first question, our hypothesis was that we would be looking at discourses about prostitutes, not at discourses arising from prostitutes themselves. On the basis of the analyses undertaken, this appears to be true. Prostitutes are the object of discussion in the EEBO corpus, but they are not given a voice. While the possibility exists that in the corpus there are a few texts in which their voice, un-mediated, survives, the overall pattern seems to be that they are spoken about – whether that is in an account of their actions or in a complaint at their existence. When exploring such historical discourses, especially about marginalised groups, we should be ever mindful that there is one voice that may well be silenced – the voices of the very people we are interested in studying, as in this study.

Regarding inter-disciplinarity, we would hope that the answer to this question is self-evident for any reader of this paper. The geographer faces an interesting challenge set by the historian and benefits from inputs from the corpus linguist, using concordancing, for example, to reduce a processing problem via concordance based geo-parsing. The historian benefits from the visual analysis provided by the geographer and takes insights into text processing and the understanding of discourse from the corpus linguist. The corpus linguist benefits from the mapping by gaining a firmer grasp on the intersection between space and discourse and gains rich explanations for the patterns observed in the corpus from the experience of the historian. This is a brief and illustrative summary of the benefits gained from our collaboration[24] – we doubt very much that this maps the limits of the benefits of such an interaction. Rather, we believe that it represents the first fruits of a rich harvest that can be reaped by such inter-disciplinary teams.

References

Primary Sources

Anon. (1668). *Gods justice against murther, or The bloudy apprentice executed. Being an exact and true relation of a bloudy murther committed by one Thomas Savage an apprentice to a vinter at the ship tavern in Ratliffe upon the maid of the house his fellow servant, being deluded thereunto by the instigations of a whore*. London, UK.

Bale, J. (1574). *The pageant of popes contayninge the lyues of all the bishops of Rome, from the beginninge of them to the yeare of Grace 1555*. London, UK.

Bunyan, J. (1680). *The life and death of Mr. Badman presented to the world in a familiar dialogue between Mr. Wiseman and Mr. Attentive*. London, UK.

Care, H. (1679). *A word in season being a parallel between the intended bloody massacre of the people of the Jews, in the reign of King Ahasuerus and the hellish powder-plot against the Protestants in the reign of King James: together with an account of some of the wicked principles and practises of the Church of Rome, demonstrated in their barbarous and cruel murders and massacres of the Protestants in the Netherlands, France, Ireland, Piedmont, the Albigenses*. London, UK.

Farquhar, G. (1699). *Love and a bottle a comedy, as it is acted at the Theatre-Royal in Drury-Lane by His Majesty's servants*. London, UK.

Floyd, J. (1612). *The ouerthrovv of the Protestants pulpit-Babels conuincing their preachers of lying & rayling, to make the Church of Rome seeme mysticall Babell.* Saint-Omer, France.

Fullwood, F. (1655). *The church-history of Britain from the birth of Jesus Christ until the year M.DC.XLVIII endeavoured by Thomas Fuller*. London, UK.

Latimer, H. (1549). *The seconde [seventh] sermon of Maister Hughe Latimer which he preached before the Kynges Maiestie [with?] in his graces palayce at Westminster, ye xv. day of Marche [-xix daye of Apryll], M.ccccc.xlix*. London, UK.

Latimer, H. (1562). *27 sermons preached by the ryght Reuerende father in God and constant matir of Iesus Christe, Maister Hugh Latimer.* London, UK.

[24] Areas such as literary research and religious studies can also benefit from this type of interaction, of course. See, for example, Donaldson, Gregory and Murrieta-Flores (2015) and Gregory et al. (2014).

Lightfoot, J. (1700). *Some genuine remains of the late pious and learned John Lightfoot, D.D. consisting of three tracts*. London, UK.

Milton, J. (1642). *Nevvs from hell, Rome and the Innes of court wherein is set forth the coppy of a letter written from the devill to the pope*. London, UK.

Oates, T. (1696). *Eikon basilikåe, or, The picture of the late King James, drawn to the life in which is made manifest, that the whole course of his life hath to this day been a continued conspiracy against the Protestant religion, laws and liberties of the three kingdoms: in a letter to himself, and humbly dedicated to the King's Most Excellent Majesty, William the Third*. London, UK.

Puritanomastix, A. (pseud.) (1642). [No title listed]. London, UK.

Rowlands, S. (1613). *A crevv of kind gossips, all met to be merrie complayning of their husbands, with their husbands ansvveres in their owne defence*. London, UK.

Ruggle, G. (1662). *Ignoramus a comedy as it was several times acted with extraordinary applause before the Majesty of King James*. London, UK.

Stubbes, P. (1583). *The Anatomie of Abuses*. London, UK.

Thomson, T. (1670). *The life of Mother Shipton a new comedy as it was acted nineteen dayes together with great applause*. London, UK.

Secondary Sources

Ackroyd, P. (2000). *London: The Biography*. London, UK: Chatto & Windus.

Arnold, C. (2011). *City of Sin: London and its Vices*. London, UK: Simon & Schuster.

Attwood, N. (2011). *The Prostitute's Body: Rewriting Prostitution in Victorian Britain*. London, UK: Pickering & Chatto.

Burford, E. J., & Wotton, J. (1995). *Private Vices – Public Virtues: Bawdry in London from Elizabethan Times to the Regency*. London, UK: Robert Hale.

Conway, A. (2010). *The Protestant Whore: Courtesan Narrative and Religious Controversy in England, 1680-1750*. Toronto, Canada: University of Toronto.

Corfield, P. (1972). A provincial capital in the late seventeenth century. In P. Clark & P. Slack (Eds.), *Crisis and Order in English Towns 1500–1700: Essays in Urban History* (pp. 263–310). London, UK: Routledge and Kegan Paul.

Cressy, D. (1977). Literacy in seventeenth-century England: More evidence. *The Journal of Interdisciplinary History*, *18*(1), 141–150.

Denlinger, E.C. (2002). The Garment and the man: Masculine desire in Harris's List of Covent-Garden ladies, 1764–1793. *Journal of the History of Sexuality*, *11*(3), 357–394.

Ditmore, M. (Ed.) (2006). *Encyclopedia of Prostitution and Sex Work, Volume 1*. Westport, CT: Greenwood.

Donaldson, C., Gregory, I., & Murrieta-Flores, P. (2015). "Mapping 'Wordsworthshire": A GIS study of literary tourism in Victorian Lakeland. *Journal of Victorian Culture, 20*(3), 287–307.

Emerson, G. (2002). *City of Sin: London in Pursuit of Pleasure*. London, UK: Carlton Books.

Green, J. (2008). *Chambers Slang Dictionary*. Edinburgh, UK: Chambers.

Green, J. (2011). *Crooked Talk: Five Hundred Years of the Language of Crime*. London, UK: Random House.

Gregory, I., Cunningham, N., Lloyd, C., Shuttleworth, I., & Ell, P. (2014). *Troubled Geographies: A Spatial History of Religion and Society in Ireland.* Bloomington, IA: Indiana University Press.

Gregory, I., & Donaldson C. (2016). Using geographical technologies to understand Lake District literature. In D. Cooper, C.D. Donaldson, & P. Murrieta-Flores (Eds.), *Literary Mapping in the Digital Age* (pp. 67–87). Abingdon, UK: Routledge.

Gregory, I., & Hardie, A. (2011). Visual GISting: Bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistics Computing, 26*(3), 297–314.

Griffiths, P. (1993). The structure of prostitution in Elizabethan London. *Community and Change, 8*(1), 39–63.

Griffiths, P. (1998). The meaning of nightwalking in early modern England. *The Seventeenth Century, 13*(2), 212–238.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the Edinburgh Geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London Series A: Mathematical Physical and Engineering Sciences, 368*(1925), 3875–3889.

Healey, J. (2010). The development of poor relief in Lancashire, c.1598–1680. *The Historical Journal, 53*(3), 551–572.

Henderson, T. (1999). *Disorderly Women in Eighteenth Century London: Prostitution and Control in the Metropolis, 1730–1830*. London, UK: Longman.

Hindle, S. (2003). "Not by bread only"? Common right, parish relief and endowed charity in a forest economy, c. 1600–1800. In S. King & A. Tomkins (Eds.), *The Poor in England 1700-1850*: *An Economy of Makeshifts* (pp. 39–75). Manchester, UK: Manchester University Press.

The Historical Thesaurus of English, version 4.21. (2017). Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk/ (last accessed August 2018).

Karras, R. M. (1996). *Common Women: Prostitution and Sexuality in Medieval England.* Oxford: Oxford University Press.

Kulldorf, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods, 26*(6), 1481–96.

Lancashire, I. (2015). *Lexicons of Early Modern English*. Toronto, Canada: University of Toronto Library and University of Toronto Press. URL: leme.library.utoronoto.ca

Linnane, F. (2003). *London: The Wicked City*. London, UK: Robson.

McEnery, T., & Baker, H. (2016). *Corpus Linguistics and Seventeenth-Century Prostitution*. London, UK: Bloomsbury.

McEnery, T., & Baker, H. (2017). The public representation of homosexual men in seventeenth-century England – A corpus based view. *Journal of Historical Sociolinguistics*, 3(2), 197–217.

McMullan, J. (1984). *The Canting Crew: London's Criminal Underworld, 1550–1700*. New Brunswick, NJ: Rutgers University Press.

Mowry, M. (2004). *The Bawdy Politic in Stuart England, 1660–1714: Political Pornography and Prostitution*. Aldershot, UK: Ashgate.

Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A., & Rayson, P. (2015). Automatically analyzing large texts in a GIS environment: The Registrar General's Reports and cholera in the 19th Century. *Transactions in GIS, 19*(2), 296–320.

Nevala, M., & Hintikka, M. (2009). Cider-wenches and high prized pin-boxes: Bawdy terminology in seventeenth- and eighteenth-century England. In R.W. McConchie, A. Honkapohja,

& J. Tyrkkö (Eds.), *Selected Proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HEL-LEX 2)* (pp. 134–152). Somerville, MA: Cascadilla Press.

Oxford University Press (2018), *The Oxford English Dictionary. OED Online.* Retrieved from http://www.oed.com (last accessed August 2018).

Pound, J.F. (Ed.) (1971). *The Norwich Census of the Poor 1570*. Norfolk, UK: Norfolk Record Society.

Quaife, G.R. (1979). *Wanton Wenches and Wayward Wives: Peasants and Illicit Sex in Early Seventeenth Century England*. London, UK: Croom Helm.

Rupp, C.J., Rayson, P., Gregory, I., Hardie, A., Joulain, A., & Hartmann, D. (2014). Dealing with heterogeneous big data when geoparsing historical corpora. In *Proceedings of the 2014 IEEE International Conference on Big Data* (pp. 80–83). Washington, DC: IEEE.

Sharpe, J.A. (1984). *Crime in Early Modern England, 1550–1750*. London, UK: Longman.

Shoemaker, R.B. (1991). *Prosecution and Punishment: Petty Crime and the Law in London and Rural Middlesex, c.1660–1725.* Cambridge, UK: Cambridge University Press.

Silvester, L. (2012). The experience of single women in early modern Norwich: “Rank beggars, gresse maydes and harlots”. In A. M. Scott (Ed.), *Experiences of Poverty in Late Medieval and Early Modern England and France* (pp. 85–106). Farnham, UK: Ashgate.

Stapleton, B. (1993). Inherited poverty and life-cycle poverty: Odiham, Hampshire, 1650–1850. *Social History, 13*(3), 339–355.

Taavitsainen, I. (2016). The case of address terms. In W. Anderson, E. Bramwell, & C. Hough (Eds.), *Mapping English Metaphor Through Time* (pp. 260–280). Oxford, UK: Oxford University Press.

Thompson, R. (1979). *Unfit for Modest Ears: A Study of Pornographic, Obscene and Bawdy Works Written or Published in England in the Second Half of the Seventeenth Century*. London, UK: Macmillan.

Tobin, R., Grover, C., Byrne, K., Reid, J., & Walsh J. (2010). Evaluation of georeferencing. *Proceedings of the 6th Workshop on Geographic Information Retrieval* (pp. 7:1–7:8). New York, NY: ACM.

Traugott, E. C. (2010). Dialogic contexts as motivations for syntactic change. In R. A. Cloutier, A. M. Hamilton-Brehm, & W. A. Kretzschmar (Eds.), *Studies in the History of the English Language V: Variation and Change in English Grammar and Lexicon* (pp. 11–27). Berlin, Germany: Mouton de Gruyter.

Walker, G. (2003). *Crime, Gender and Social Order in Early Modern England*. Cambridge, UK: Cambridge University Press.

Williams, S. (2011). *Poverty, Gender and Life-Cycle under the English Poor Law, 1760–1834.* Woodbridge, UK: Boydell and Brewer.

Wilson, D. (2004). *All the King’s Women: Love, Sex and Politics in the Life of Charles II*. London, UK: Pimlico.

Wrightson, K. (1973). *The Puritan Reformation of Manners*. Doctoral dissertation, Cambridge University, Cambridge, UK.

Wolfgang Teubert

Corpus linguistics: Widening the remit

Abstract: Language allows us to turn our experiences into meaning and share them with others. This is why linguistics, and corpus linguistics in particular, should have a strong focus on the meaning of what is said. If we accept the meaning of a lexical item such as *human rights* to be everything said about it (the paraphrastic content of all occurrences of this item), we'll find that the Firthian concept of meaning (1957: 196) as an "abstraction on the syntagmatic level", which has been the core of corpus-oriented collocation studies, does not really allow us to make sense of what has been said. Statistics is no more than a heuristic tool; it makes us aware of what may be relevant. The methodology of corpus linguistics must include paraphrase analysis. The aim is to extract, organise and present the relevant textual evidence. It is then up to the discourse participants themselves to interpret these findings and come up with new paraphrases.

1 Some comments on the study of meaning

As a linguist, I am interested in meaning more than in anything else. There are many other fields, for instance the study of the language system with its universal laws and arbitrary rules, the study of pragmatics, aimed at explaining how we understand each other beyond the confines of the language system, and the study of language variation and of language change. Corpus linguistics proper, if there is such a thing, is less concerned with theoretical linguistics, the playground of cognitive linguists and of those speculating about the biological foundation of language, among others. Still if we look at our corpus journals we notice corpus linguistics covers most of the other areas mentioned above, even if it replaces strict formal rules by the statistics of the (im)probability of identifiable surface elements co-occurring. However, looking at recent volumes of corpus journals, it seems to me the study of meaning plays a lesser role.

I very much regret that the contribution corpus linguistics has made to the study of meaning is not yet as exciting as it could be. For it is meaning that makes language special, compared to non-symbolic human interaction (collective singing, hunting, fighting). Arbitrary signs, the essence of all language, allow us to

Wolfgang Teubert, University of Birmingham, w.teubert@bham.ac.uk

https://doi.org/10.1515/9783110489071-006

share and exchange information, knowledge, ideas, and that enable us to co-operate in collectively planned activity. I find it therefore disheartening that there are introductions to corpus linguistics which do not feature meaning at all, for instance Graeme Kennedy's *An Introduction to Corpus Linguistics* (1998). Meaning is also what distinguishes language from similar edifices such as mathematics or music. We talk not so much to construct a novel grammatical sentence, in the way new equations are solved, and also not to entertain others through the melodious prosody of our voice, but to make sense of ourselves and the world around us. It is the inherent dialogic nature of language that not only allows us to share and exchange symbolic content, but also stimulates new ideas by exchanging and sharing symbolic content with people around us.

Corpus linguistics is proud of its strict scientific methodology. But should we really look at language studies as a science on par with the natural sciences? Is any given language, any discourse, not also a cultural achievement in need of interpretation? Many linguists see themselves as scientists. For cognitive linguists, there is little doubt that their remit is part of the cognitive sciences, the more recent biolinguistic enterprise views itself as an extension of evolutionary biology, neurolinguistics carries out experiments in labs using highly complex apparatuses, and empiricist corpus linguistics has developed a plethora of tools that applied to the same corpus invariably bring about the same results. But their findings do not tell us what a text, or a text segment, maybe just a word, means. It is true that year by year the summarisers developed by computational linguists come up with more useful results, and translation systems have improved vastly over past decades. All these disciplines, as well as including, until very recently, corpus linguistics as well, keep their distance from the study of discourse. They are not keen to interpret a haiku or tell us what Vladimir Putin really wanted to say when he gave his 2014 Sochi speech. However, things are gradually moving. Increasingly, linguists of many backgrounds and even more so social science scholars analysing discourse have turned to corpus linguistics, encouraged by books like Paul Baker's (2006) *Using Corpora in Discourse Analysis*. By comparing a single text or a collection of texts to a much larger reference corpus, corpus tools indeed can help find patterns and striking correspondences (in the form of intertextual links) normally overlooked by human readers, aiding the interpretation of the text segment, the text or the text collection in question. But while the appropriate tools deliver results, they do not provide an interpretation. When it comes to meaning, the results they deliver are useful inasmuch as they help a participant in such a discourse or a researcher make sense of the expressions they encounter.

2 Some other ways to look at meaning

What is meaning? This question is far from new. Our lexicographical traditions let us think, first of all, of the meaning of words. Unsurprisingly, many different things are said about word meanings, as a brief Google query shows:

(1) It has also been suggested that **the meaning of a word** is simply the entity in the World which that word refers to.[1]

(2) **The meaning of a word** [e.g. *rose*] is a link to an entry in the person's mental encyclopedia, which captures the concept of a rose.[2]

(3) This theory [of lexical semantics (D.A. Cruse)] understands that **the meaning of a word is** fully reflected by its context. Here, the meaning of a word is constituted by its contextual relations.[3]

(4) The hypothesis I propose is that the following is an acceptably Humean view: … **The meaning of a word is the** custom, connected with the word, that enables us to pass from one idea correlated with the word to other ideas that resemble it. Understanding the meaning of a word would consist in the acquirement or possession of such a custom, not in merely possessing or forming an idea, in connection with the word.[4]

Since ancient times there have been two main traditions, one in which the meaning of a lexical item is its reference to the discourse-external, real-world kind of thing it stands for, and the one in which the meanings of expressions refer to mental concepts into which our perceptions of the world are automatically translated. But does it make sense to situate a thing like *human rights* in the real world, outside of discourse? That words refer to mental concepts, understood as ideas, is a more popular conception. One strand of this tradition, from Aristotle to the Jerry Fodor of *The Language of Thought* (Fodor 1975), argues that these concepts are innate, immutable and universal. Stephen Pinker allows, in the second quotation, for the acquisition of new concepts into the encyclopaedia people have in their heads. The third quotation refers to Alan Cruse's theory of word meaning and would be fully endorsed by corpus linguists. 'Context' is, of course, a fuzzy

[1] Retrieved from https://sites.google.com/a/sheffield.ac.uk/all-about-linguistics/branches/semantics/what-is-semantics (last accessed October 2018)

[2] Retrieved from http://www.worldcat.org/wcpa/servlet/DCARead?standardNo=0465072690&standardNoType=1&excerpt=true (last accessed October 2018)

[3] Retrieved from http://en.wikipedia.org/wiki/Semantics (last accessed October 2018)

[4] Retrieved from https://www.jstor.org/stable/2107059 (last accessed October 2018)

word itself, and sometimes, particularly in pragmatics, seems to mean not just the text around it (John Sinclair's 'cotext', is often in corpus studies limited to the narrow window of -4/+4 words around the node) but also the wider situation into which an utterance is embedded (even going beyond what is actually said). I sometimes wonder why my own view, namely that the meaning of a lexical item is what people have assigned to it, by agreeing (and also by disagreeing) about what it should mean, is shared less by other linguists than by philosophers such as David Hume (see last quote above), Friedrich Nietzsche and Jacques Derrida.

3 Meaning and interpretation

The overriding problem in much of the discussion of 'meaning' is that meaning often is not properly distinguished from interpretation. When talking about the meaning of a word it is best to exclude mentalist and cognitive approaches, because we cannot look into people's heads. Meaning is found in discourse, in the form of textual evidence. The meaning of an expression is what has been said about this expression (or about the parts of which it consists). People define and use expressions in different ways, and when expressions are controversial, there often is no common denominator for what is said about them (cf. Chapter 16 in Teubert 2010). It is the sum of the negotiations concerning this expression, consisting of the paraphrases that people come up with in their endeavour to make (their own) sense of them, i.e. to interpret the textual evidence available to them of a given expression. Thus, from a diachronic perspective, we find layer upon layer of interpretation of a given expression, as each use of it embeds it in a context providing paraphrastic content. Some paraphrases may be more frequent than others, some may occur more commonly in a certain network of texts than in other networks, and some older ones may become rarer and some new ones may become more frequent. Taken together, they constitute the meaning of the expression. There is no right or wrong paraphrase. The successful ones leave traces in subsequent utterances, while the others disappear. The meaning of a lexical item or of an expression exists solely in what has been said about it, and not in people's heads.

Thus, the meaning of a lexical item or a larger expression or text segment is what has been said about it, is the textual evidence we have for it. To make sense of it, this evidence has to be interpreted, to be paraphrased. And each new paraphrase adds something to the meaning. Interpretations are more than summaries of the textual evidence. An interpretation is a creative act. Its outcome is not determined. We don't know why people say what they say. But as they want to be successful, they will normally say it in such a way that at least some in their audience will accept it. They want other people to refer to it, to repeat it, or to comment on it, or even to reject it. They want to make an impact on discourse. Thus they will want to balance the novelty of their paraphrase with the continuity of what has been said previously.

Taken together, the paraphrases we find for a given lexical item or expression are testimony of the ways in which people negotiate their meaning. Stanley Fish calls these situations in which people exchange and share their interpretations 'interpretive communities' (cf. Fish 1980). There is no need for those taking part to agree on a single interpretation. It is no more than an exchange of opinions. Even after long discussions, it is common for people not to agree.

Thus there are two perspectives. One focuses on discourse, taking into account its diachronic dimension, regarding every new contribution featuring a given lexical item in its paraphrastic context as a reaction to previous such contributions. It studies the intertextual links by which these contributions are connected. It does not deal with the speaking subjects and with what makes them say what they say. The other perspective focuses on ourselves, on the interpreting subjects (on the 'readers' in the sense Roland Barthes uses this term). When we, in order to make sense of a given lexical item, acquaint ourselves with (some of) the apposite paraphrastic evidence extracted from discourse, we interpret this evidence, convinced we, as intentional subjects, understand what we find there and can draw our conclusions, in the wider context of all the symbolic content of which we are aware. We enter our own interpretation into discourse, thus providing yet another paraphrase for this lexical item, in the hope it will make an impact on the way it is used in subsequent contributions. For others, for those who are taking stock of our interpretations, what we have said will become part of the meaning of this lexical item. Thus we can view, along with Jacques Derrida in *Limited Inc* (Derrida 1977), the iteration of occurrences of a given lexical item as the sequence of its interpretations.

We also have to put aside the claim that the meaning of a lexical item is part of or determined by a language system, whether it would be, in the Chomskyan version, universal and innate or, in the Saussurean version, language-specific, in any case a system on which whatever is said would have no effect. It is people who collectively engage in meaning-making, and every new paraphrase of a lexical item, every new interpretation, arbitrary as they will be, makes an impact on its meaning, as long as subsequent contributions refer to it.

Corpus linguistics is irreplaceable in detecting lexical items, sorting them in terms of their collocates and other patterns of usual usage, thus disambiguating different senses, and perhaps, in future developments, discovering intertextual links between citations. Its scientific methodology delivers results, informing us, among many other things, of raw and relative frequencies of lexical items and of recurrent paraphrases including their variations. What it does not do is to venture an interpretation of its findings. Interpretation is what only people can do. This is why we cannot expect to develop a scientific methodology, translatable into software tools, that would make sense of a lexical item or a text segment. It is this gap between statistical evidence and human interpretation that even Patrick Hanks, acclaimed corpus linguist and lexicographer, finds impossible to bridge (Hanks 2013; Teubert 2016).

4 The quest for meaning: The enigma of collocation

In this part, I will take a closer look at the 'scientific methodology' of corpus linguistics, which is often seen as the key attraction of the corpus-driven approach, putting it seemingly on par with the natural sciences. It is a view shared, for instance, by Tony McEnery and Andrew Hardie: "the combination of falsifiability and replication can make us increasingly confident in the validity of corpus linguistics as an empirical, scientific enterprise" (2012: 16). To which extent does the automatic, and thus unbiased, application of analytical tools on a corpus help us in our quest for meaning? The example I will discuss is the lexical item *human right*, or, occurring more frequently, its plural form, *human rights*. Because it is such a controversial notion we will find an abundance of paraphrases, and this makes it a suitable example for my purpose. The key concepts of corpus linguistics are the notion of collocation and patterns, the role of intuition vs. statistics, the concept of probability, to which I believe should be added paraphrastic content (as distinguished from the insufficiently defined concept of 'use') and finally intertextuality and the diachronic dimension of discourse. I would like to begin by showing that, because there is no common denominator for the meaning of *human rights*, a straightforward algorithmic generalisation of meaning cannot work. My focus here is the universality of human rights, seen by many as an essential property of this concept and thus part of its core meaning. The following citations are extracted from the web, using Google. The queries in this paper were carried out between November and December 2014, and the query phrase is in bold; the numbers in brackets give the frequency of the phrase.

(5) **Human rights are universal** (290) **because** they are based on humanity as the sole shared aspect in a world of different nations, cultures, religions and traditions.

(6) **Human rights are not universal** (154) **because** different people in different countries have different sets of rules and regulations.

(7) Human rights are held by all persons equally, universally, and forever. **Human rights are inalienable** (297)**:** you cannot lose these rights any more than you can cease being a human being.

(8) **Human rights are not inalienable** (23) or universal, but must reflect the cultural and other backgrounds of states and individuals.

Judging from the immense social studies literature on human rights, the adjective phrases *universal* / *not universal* and *inalienable* / *not inalienable* are, it seems, relevant (if not very frequent) collocates. As we will see, they are important constituents of the meaning of *human rights*, in spite (or perhaps because) of them being contradictory and thus constantly under negotiation. But how relevant are they from the Firthian perspective of meaning by collocation? Even though the role of collocation for the notion of meaning was already recognised by Harold Palmer before the Second World War (Palmer 1933), it was J. R. Firth with whom it is now commonly associated. This is his ubiquitously quoted definition:

> Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*. (Firth 1957: 196)

It is quite evident that for Firth collocation has nothing to do with what I call the paraphrastic content we find associated with a lexical item. Collocation in his sense is not about the paraphrases people have entered into discourse (i.e. have offered to an audience) and not about the "conceptual or idea approach to the meaning of words" (1957: 196). One might not be entirely wrong to translate "abstraction at the syntagmatic level" into 'a formalised representation of a syntactic structure', thus moving his definition closer to what is now generally called colligation. As he sees it, meaning is not something like the gloss we find in a dictionary entry, describing what a given word is about. Perhaps we would be right to assume that for him the consolidated meaning of a node such as *night* is a formula gained from calculating the probabilities of finding collocates in syntagmatically defined positions in the narrow context of the node, while each collocate, in this case *dark*, (and similarly each combination of collocates) in conjunction with the node, provides one of the (rather large number of) meaning formulae of the node, equivalent to the number of diverse contexts in which the node is embedded.

> In short, the term collocation denotes the idea that important aspects of the meaning of a word (or another linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words and or structures with which it frequently co-occurs […]. (McEnery & Hardie 2012: 122–123)

Tony McEnery and Andrew Hardie paraphrase Firth by saying that collocation is the notion that the meaning of a lexical item is largely defined by the "characteristic associations" the node has with other words around it, in relationship with the (syntagmatic?) structures obtaining between node and collocate, only to add that nowadays "we find a great multitude of definitions" (2012: 123). They point out that in many of them intuition (and thus the "conceptual or idea approach" rejected by Firth) is foregrounded. McEnery and Hardie, however, side with Firth and therefore "will not consider any word co-occurrence identified by researcher intuition" (2012: 123). What remains are co-occurrence patterns based on raw frequency or on statistical significance, in the way of (sometimes discontinuous) n-grams or lexical bundles. The authors draw attention to the "problems inherent in determining collocations" and the "issue of collocation statistics". In summing up the current state, they see the "emergence of at least two distinct schools of thought in neo-Firthian linguistics", collocation either considered, in line with Michael Hoey, as "a mind-internal phenomenon", or, in line with Teubert, as "a tool for exploring discourse" (McEnery & Hardie 2012: 122–133). In my view, the authors' dismissal of the role of intuition ("the conceptual or idea approach") is somewhat premature, even if essential for claiming scientificity.

What McEnery and Hardie call intuition is, as I hope to show below, an important part of approaching meaning. Linguists (like researchers in all other sciences) are constantly required to rely on their experience, on what they have learnt so far. Lexicographers, too, even those in corpus linguistics, rely on it, as no one has come up with a recipe to automatically generate formulae expressing the meaning of words to the satisfaction of a normal language user. For to assemble, order and present the textual evidence for someone who wants to know what a lexical item means it is necessary to make any amount of arbitrary decisions, as we can see if we compare the entries of different dictionaries. Any lexicographer's interpretation cannot be but biased. But computers cannot do it. There is no computer program that would tell a language user the meaning of *human rights*, not even in form of an "abstraction at the syntagmatic level", much less as a verbal formulation representing "the conceptual or idea" meaning. As long as people are free to use words however they like, there is no mathematical trick that could possibly find a common denominator for how they have paraphrased the meaning of the lexical item in question.

Outside of computational linguistics, there is no one who has taken Firth further than Bill Louw. He, too, makes short shrift of intuition because, he says, it introduces bias where science and automated processes should rule. He is convinced that there is some universal logic, some immutable system underlying language, and once we achieve to model it, it will uncover what he calls the subtext, the true meaning hidden behind the superficial ideas and concepts of which a text seems to speak.

> The use of ‘prompts’ from philosophy that are capable of verifying meaning through collocation within natural language provide us with highly innovative methods of investigation. This technique involves the provision of increasingly strenuous Popperian tests (1) that are born of alterations to normal investigative procedure; (2) an entry point that is entirely shielded from bias because it operates with and through human intuitive opacity to the logic of natural language, (3) thus yielding insights into meaning that cannot be obtained through intuitive methods that have always dominated the discipline. (Louw 2014: 7)

Louw’s valuable contributions to linguistics, his conceptualisation of semantic prosody and his work on irony, for instance, show, I believe, very much the creativity of a romantic genius driven by his intuitions, leading him here and there. Yet I seriously doubt that handing over the messy issue of coming to terms with lexical meaning to automated processes really delivers results which we do not have to submit to our scrutiny and interpretation. Anyway, as I see it, the point of uncovering a subtext in order to understand the idea expressed in a text, for instance in a poem, is part of an interpretive act, driven very much by intuition.

Firth has never spelled out of what kind of an abstraction exactly the “abstraction at the syntagmatic level” is (1957: 196). Do we have to think of established categories such as the parts of speech, semantic attributes (animate, human, ideational etc.), syntactic roles (noun phrase, prepositional object etc.), or will these syntagmatic features only reveal themselves by our investigation of the collocational nature of language? This is, according to McEnery and Hardie (2012), what distinguishes the Birmingham school, with John Sinclair, the doyen of corpus linguistics, from the Lancaster one, with Geoffrey Leech as its originator. This division is also categorised by the opposition between the corpus-driven and the corpus-based approach or, as the authors prefer to label it, between corpus-linguistics-as-theory and corpus-linguistics-as-method.

It was Elena Tognini-Bonelli who introduced the corpus-driven approach to a wider public. The two relevant sentences read:

> The corpus-driven approach builds up the theory step by step *in the presence of the evidence*, the observation of certain patterns leads to a hypothesis, which in turns leads to the generalisation in terms of rules of usage and finally finds unification in a theoretical statement. (Tognini-Bonelli 2001: 17)
> The theory has no independent existence from the evidence and the general methodological path is clear: observation leads to hypothesis leads to generalisation leads to unification in theoretical statement. (Tognini-Bonelli 2001: 85)

The attraction of the corpus-driven programme is its claim to a theory borne out of an empiricist methodology without any preconditions, in other words, pure science. The starting point is a tabula rasa. All categories and all relationships obtaining between them will result from nothing but observation, from the application of tools operating automatically, without human interference or bias. Analysing real language data will deliver patterns of co-occurrence between the elements of which language consists. Traditional categories may have been useful for describing Latin; but are they also appropriate for more isolating languages such as English or Chinese? It is an approach that makes us reconsider received categories like the parts of speech; it will give us a perspective of language as it 'really' is. Of course, Tognini-Bonelli was as aware as Sinclair that this corpus-driven programme can be no more than a utopian goal. We have to read it as a call for more awareness that by annotating a corpus, for instance by tagging certain words as adjectives, a subsequent analysis of such tagged adjectives won't tell us anything new about the category 'adjective'. The corpus-driven approach sees itself as basic research, promising to deliver new insights about language, while the corpus-based approach is more useful for developing language technology, based on what we already know about language, in the context of computational linguistics or language engineering.

Ten years ago, I, too, praised the corpus-driven approach in my paper 'My version of corpus linguistics'. However, in applying it to my programme, I slightly altered its direction:

> While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach. (Teubert 2005: 4)
> This is what makes paraphrases so essential: they tell us what has been said and can be said about a discourse object. For a corpus-driven theory of meaning, they are crucial. They may contradict each other, they may describe something in such irreconcilable features that it is hard to see it as the same thing, but taken together in all their chaotic diversity they are the very material meaning consists of. (Teubert 2005: 12)

The categories defining the system of mechanics can be discovered from the outside. The categories defining a given language, on the other hand, have been invented, by the people using this language. By talking about language they construct it, along the lines they see fit. Mothers train infants to speak properly. A language does not have adjectives before people start talking about them. In the case of oral communities, these people are in our days normally anthropological linguists. Their descriptions will impact the kind of normalisation required for literate languages.

Human language is not a natural kind. It is a cultural artefact. There is no language system unless language users make it. Homo sapiens had been around for over a million years before people began talking to each other. Homo loquens came late, perhaps only one or two hundred thousand years ago. We have to view language as a cultural achievement, a creation similar to the one of the first cave paintings, which also appeared in the middle paleolithic era, somewhere between 200,000 and 40,000 years ago. Scepticism is called for when claims are made about language universals which seem to point to some innate language faculty. Nicholas Evans and Stephen Levinson sum up their critical stance:

> Talk of linguistic universals has given cognitive scientists the impression that languages are all built to a common pattern. In fact, there are vanishingly few universals of language in the direct sense that all languages exhibit them. Instead, diversity can be found at almost every level of linguistic organization. (Evans & Levinson 2009: 429)

Meaning does not reside in the neurons and synaptic connections of our brain. There is no universal language of thought. The meaning of a lexical item is what people have said about it. This is what linguists have to study. There is no discourse-external taxonomy to which we can take refuge. All the categories applying to language are discourse constructs; and it is these categories that have moulded what we call our language. The discourse of linguists is part of this ubiquitous discourse, just as the western notion of science and objectivity is inseparable from the western culture that has developed it. Linguistic categories are neither immanent nor immutable; people, including linguists, will continue to renegotiate them as they have done over thousands of years.

Collocation is such a category. Today practically all linguists agree that the co-occurrence patterns they reveal shed light on meaning. But while some see the measurement of collocation as a discovery procedure alerting us to paraphrastic content we might otherwise have overlooked, it is, for others, a deeper, more abstract and therefore unbiased notion of meaning. Yet for gauging the significance of collocates we have to choose among a wide spectrum of statistical operations which, automatic as they are, deliver a diversity of results none of which can claim objectivity. Of the dozen or so statistical procedures available for collocation analysis, the linguist normally chooses the one which they think will give them the results most in line with their research question. Using Mutual Information (MI, left) and t-score_(right), I obtain these results for the occurrences of *human rights* in the Bank of English:

Tab. 1: Collocates of *human rights* in the Bank of English.

MI	t-score
Zoeg (3)	Human (18496)
Minkahyup (3)	European (1441)
Abdennour (4)	Abuse (1194)
Haldi (3)	International (982)
Jumale (3)	Commission (824)
Rafto (3)	Of (5872)
Tapol (5)	Violations (722)
Komnas (4)	Court (778)
Yayasan (9)	Groups (740)
Demba (3)	Record (685)

It would be easy to dismiss the MI results. These strange words, names, actually, are found rarely in our media discourse, which accounts for a large part of the Bank of English. MI typically measures the degree in which a co-occurrence of two items defies statistical probability and thus has a bias for rare items, while t-score foregrounds co-occurrence patterns based on the high frequency of the collocate. The MI results would be meaningful for a researcher interested in lesser known human rights campaigners. T-score, on the other hand, works well to provide evidence for what we already believe to be the case, for what corresponds to our expectations, while also pointing out some co-occurrences we were not aware of. (Interestingly, neither *universal* nor *inalienable* come up among the ten most significant collocates, indicating that they belong to more academic discussions of human rights.) Human rights commissions and similar groups on the international level and the European Court of Human Rights look at countries' human rights records and find instances of *abuse* and *violations*. We are not surprised; abuse and violations abound when human rights are discussed. But in the case of *abuse*, my expectations have fooled me. For I was thinking of citations like this (search phrase: *human rights* + *abuse*):

(9) *Telegraph, 16 Nov 2014* China is guilty of “large scale and systematic” **human rights abuse**s, the Deputy PM has said on the first morning of a three-day visit to London by Premier Li Keqiang.

Actually, citations like this are relatively rare. This is why corpus linguistics is essential in any quest for meaning. Its tools provide findings which work as a check for our prejudices, our preconceived ideas. Typically, *abuse* co-occurs with *woman*, *girl*, *daughter* or *child*, in nine out of the first ten hits (Google) for ‘he abused’. But in the large majority of *abuse* as a collocate of *human rights*, it is, at least in British media discourse, human rights legislation which is abused, namely by criminals and terrorists (Google):

(10) *Scottish Daily Record-13 Nov 2013* … the **Human Rights** Act as it stands is not fit for purpose and too open to **abuse**.

(11) *Yorkshire Post-18 Nov 2013* This is all on top of the **Human Rights** Act which incorporated the [European Convention of Human Rights] **…** We, in this country, do not need lectures on real human rights and if the killers, terrorists and people trying to **abuse** our country and its tradition of fairness do not like it – as far as I am concerned, that is just tough.

(12) *Daily Mail-7 Nov 2013* The **abuse** of the **human rights** legislation has now reached such offensive proportions that I think we have to consider repealing it in this Parliament, [Sir Gerard Howarth] said.

(13) *Telegraph.co.uk-22 Nov 2013* … the **Human Rights** Act and the European Convention on Human Rights and their apparent **abuse** by criminals and terrorists.

McEnery and Hardie distinguish two ways to study collocation. One is what they call ‘collocation-via concordance’; described as “the linguist’s intuitive scanning of the concordance lines that yields up notable examples and patterns, not an algorithm or a recoverable procedure” (2012: 126). Whoever remembers John Sinclair’s lectures knows that he often did just that: he took his audience through a large number of concordance lines online delivered from the Bank of English, showing the patterns in which words like *budge* were found quite regularly. What McEnery and Hardie call intuition is reminiscent of his corpus-based approach. They regret that collocation-via-concordance, i.e. “the linguist’s intuitions and hand-and-eye methods”, take “for most neo-Firthians” precedence over “the utility of statistics” (2012: 127). ‘Collocation-via-significance’, on the other hand, is when “statistical tests should have the key role in determining collocation” (2012: 127). It also has problems, the authors inform us, in that there are a variety of statistics programs each measuring significance differently.

However the underlying issue is a different one. The problem is that the discussions of collocation often do not distinguish between meaning and interpretation. The meaning of a lexical item is the entirety of what has been said about it, or about the discursively constructed object for which this lexical item stands. To make sense of this entirety (or the sample available to us) we have to reduce the complexity of this textual evidence. We have to replace discourse in general (everything said anytime anywhere) by a special discourse we have defined, in line with our research question, in such a narrow way, that we can decide for each text whether it is part of this discourse or not. For instance, we could define the discourse for our investigation as all the texts in British national mainstream media dealing with human rights over the last ten years. We have to limit the context around each occurrence of *human right(s)* we take into account, perhaps not to the -4/+4 window most corpus tools use but to the sentences around the sentence in which the item occurs. This is what gives us the meaning in form of paraphrastic content. But we have to be aware that what comes up as paraphrastic content is determined by arbitrary decisions, such as the corpus we compile, the width of context we take into account, and, last but not least, what is counted as paraphrastic content and what is not. The researcher's decisions impact on the subsequent interpretation of the evidence. Therefore the result cannot be free from bias. Neither a participant in this human rights discourse nor the researcher observing them can claim their interpretation to be final, conclusive or 'true'.

The idea of a scientific, corpus-driven approach is also in conflict with the necessity of defining a research question guiding us through an enquiry. We must begin with an idea of what we are looking for, preliminary as it may be. This research question will have to be expressed in categories predating the enquiry, and thus our approach can never be strictly bottom-up. Any research question will, of necessity, introduce bias into our quest for meaning. From all the statistical tools, the researcher will choose the one promising to deliver, better than the others, the kind of collocates they are interested in. They all are programmed to distinguish between what is 'significant' and what is not. It is, however, worth looking at the notion of 'statistical significance'. Google points us to this definition which is just about simple enough for me to make sense of it:

> All results obtained by statistical methods suffer from the disadvantage that they might have been caused by pure statistical accident. The level of statistical significance is determined by the probability that this has not, in fact, happened. P is an estimate of the probability that the result has occurred by statistical accident. Therefore a large value of P represents a small level of statistical significance and vice versa. (Numberwatch n.d.)

For me, a keyword here is ‘probability’. Statistical significance is modelled to reflect the probability of something happening. What corpus linguists are after are not random co-occurrence patterns, but those defying probability. That is what ‘significance’ means. Yet how real is probability? Is it more than a figment of our collective imagination? Would there be probabilities if no one talked about them? In Maria Carla Galavotti’s paper ‘Subjectivism, objectivism and objectivity in Bruno de Finetti’s Bayesianism’ we read:

> For de Finetti, probability is always subjective and expresses the degree of belief of the evaluating subject. His perspective does not accommodate a notion of “objective chance” …. To de Finetti’s eyes, objectivism, namely the idea that probability depends entirely on some aspects of reality, is a distortion, and the same holds for the idea that there exists an absolute notion of objectivity, to be grounded on objective facts (Galavotti 2001: 165).

Bruno de Finetti is generally regarded as the foremost Italian mathematician of the 20th century, and he is credited with the development of the subjectivist school of probability. His views are widely accepted today. His stance accounts for the existence of a wide range of statistical tools from which corpus linguists can choose in order to obtain the evidence they are looking for, reflecting their naturally biased expectations (‘previsions’, as de Finetti calls them). Perhaps this is why we read in *Collocation: Applications and Implications* by Geoff Barnbrook, Oliver Mason and Ramesh Krishnamurthy this remarkable statement:

> It is difficult to integrate the operation of collocation fully into linguistic theory… The lack of a fixed relationship between a word and its collocate, other than a mere co-occurrence within a text, means that collocation is overall of limited value to linguistic theory. (2013: 172)

Perhaps, then, we have to accept that the phenomenon of collocation, the backbone of corpus linguistics, is, applied to the quest for the “concept or idea” aspect of meaning, most useful as a discovery procedure, and not as its replacement.

5 Meaning by collocation: the syntagmatic level

What can we learn from studying the syntagmatic level of collocation? How, for instance, can we distinguish *human rights* from *fundamental rights* and from *civil rights*? Here is a small random sample, taken from the British National Corpus (BNC):

(14) the civil rights movement
(15) the pre-eminent champion of civil rights
(16) the Derry-born writer and civil rights activist
(17) the Paisleyites blocking the civil rights marchers sayable
(18) a wholesale violation of civil rights
(19) the civil rights rally
(20) ‘Civil Rights March’
(21) the blue civil rights banner
(22) fundamental rights of citizenship
(23) parents' fundamental rights
(24) violations of fundamental rights
(25) fight for the fundamental rights of man
(26) the fundamental rights enshrined in the Constitution of the Republic of Namibia
(27) the non-violent exercise of their fundamental rights
(28) constitutional instruments establishing fundamental rights
(29) a recent meeting in London on human rights in Sri Lanka
(30) UN Commission on Human Rights
(31) a French human rights group
(32) one of the basic fundamental human rights
(33) the prominent Slovak lawyer and human rights activist
(34) the European Convention on Human Rights
(35) Britain Sikh Human Rights Movement
(36) a violation of human rights
(37) the European Convention on Human Rights

Noun phrases like these, with the lexical items they contain, seem to correspond to Firth’s phrase of ‘collocation at the syntagmatic level’. Reading these lines, we learn very little about the ‘concepts’ or ‘ideas’ for which the items in question stand. Uninterpreted, they do not help us distinguish *fundamental rights* from *civil rights* or from *human rights*. Firthian ‘meaning by collocation’ here is perhaps a cipher for what is sayable, and it is sayable because it has indeed been said. But this evidence does not automatically render the kind of explanations of the meanings of lexical items users expect to find in the dictionaries they consult. They are looking for an interpretation of the evidence. Collocation at the syntagmatic level is important because it can help lexicographers include, in their glosses, key collocates and recurrent syntagms and point out what might have been overlooked. But to this day it is the lexicographers who, informed by the findings generated by collocation software, will provide the interpretation (the glosses), taking into account also the wider contexts of syntagmatically defined phrases. Such an interpretation is the result of having made decisions each of which could also have gone the other way.

The output generated by collocation software is useful in another way, too. Teachers can advise learners of English to use phrases like these if they want to emulate native speakers. There are, however, some drawbacks in reducing meaning to such syntagms. We learn there are *fundamental human rights*, even *basic fundamental human rights*, while there is no evidence in the BNC for *fundamental rights abuses* or of a *fundamental rights movement*. We find the complete set *violation of civil/fundamental/human rights* and also instances of *civil rights activist and of human rights activist*, but none of *fundamental rights activist*. Google, however, also lists *fundamental rights activist*. It is rather obvious that corpus evidence of "collocation at the syntagmatic level" is a fairly random selection even in large corpora. Most of what has been said orally has never been recorded outside of (hardly reliable) human memories. And we have to bear in mind that what is sayable is always much more than what has been said. Otherwise languages would not change. On the other hand, looking at the wider contexts of these citations, the 1993 British discourse recorded in the BNC displays trends which are perhaps less interesting to language learners but certainly noteworthy to social studies researchers. The phrase *civil rights* in the BNC refers markedly often to the Northern Ireland troubles, while *fundamental rights* seems to be linked to legal or constitutional issues, often in the context of third world countries. *Human rights*, finally, has a more European or British ring to it. Such observations go beyond both the traditional notion of lexical meaning and that of the Firthian abstractions at the syntagmatic level; yet they may be welcome, I believe, by people looking up lexical items of this kind.

6 Moving on: from collocation to paraphrastic content

Tools delivering collocation profiles for lexical items or other expressions work automatically and therefore are relatively free from human bias, even if it is a human researcher who has compiled the corpus and purposefully chosen the tool in question so that it comes close to their expectations. Collocation profiles also point us to relevant words we did not think of. However, as said above, they do not easily translate into a gloss, an interpretation, or a paraphrase of the node item. It takes an interpretive effort, something only people participating in a discourse or linguists as discourse observers, but not machines, can do. It also requires a focus determined by the research question. For a lexicographer, it may be the compilation of a dictionary entry answering the questions dictionary users would most probably have when looking up this item. This could include also some encyclopaedic meaning, i.e. what has been said about the discourse object for which this lexical item stands. In the case of human rights, it would take account of those arguments we find pro and con the universality and inalienability of human rights. The United Nations Population Fund provides this seemingly comprehensive pro argument (here, and in the following, I have underlined what I take to be key expressions; the search phrase is in bold):

(38) **Human rights** are universal and inalienable; indivisible; interdependent and interrelated. They **are universal because** everyone is born with and possesses the same rights, regardless of where they live, their gender or race, or their religious, cultural or ethnic background. Inalienable because people's rights can never be taken away. Indivisible and interdependent because all rights – political, civil, social, cultural and economic – are equal in importance and none can be fully enjoyed without the others. They apply to all equally, and all have the right to participate in decisions that affect their lives. They are upheld by the rule of law and strengthened through legitimate claims for duty-bearers to be accountable to international standards.[5]

Have they forgotten something? Here are some other arguments we find, many of them of a somewhat tautological nature, e.g.:

(39) **Human rights are universal because** they apply to everyone in the world irrespective of caste, creed or colour.[6]

(40) **Human rights are universal because** they are inherent in human beings everywhere. They are part of our humanness.[7]

[5] Retrieved from http://www.unfpa.org/resources/human-rights-principles (last accessed October 2018)

[6] Retrieved from https://books.google.co.uk/books?id=jvKOM2MHm-kC&lpg=PP1&dq=isbn%3A8180696790&pg=PA46#v=onepage&q&f=false (last accessed October 2018)

[7] Retrieved from https://www.bushcenter.org/publications/articles/2014/11/the-bush-institute-talks-with-former-chair-of-un-commission-of-inquiry-on-human-rights-in-north-korea.html (last accessed October 2018)

(41) **Human rights are universal because** they apply to everyone in the world. All human rights belong to all people.[8]

Other arguments invoke aspects like dignity or equality:

(42) **Human rights are universal because** they are based on every human being's dignity[9]

(43) **human rights are universal because** they arise out of the common equality of all persons[10]

Google also provides evidence for the opposite stance, namely that human rights are not universal. There are 19 occurrences of the phrase *human rights are not universal because*. Again we find numerous mainly academic text segments featuring rather sophisticated arguments, for instance in this PhD dissertation, which regrets that the Universal Declaration of Human Rights is not grounded in natural law but the result of arbitrary decisions of human beings:

(44) The Universal Declaration of **Human Rights**, written in the wake of World War II, was meant to provide a moral standard for judging the state's treatment of the individual. Yet to this day some contend that the principles expressed therein **are not universal**, but culturally relative. The dominant arguments for universality, however, are themselves relativistic **because** they are not grounded in the idea of a natural order that supplies objective standards of value.[11]

That human rights are a western concept based on the contingent notion of individualism and other cultural idiosyncrasies, is a common argument:

(45) Currently there have been various debates both from scholars and government that **human rights are not universal** but that cultural diversity influences what obtains as human rights in non-western states. They argue

[8] Retrieved from https://web.archive.org/web/20180112163316/http://www.peopleswatch.org/dm-documents/Resource_Material_Training_on_Human_Rights_to_Professional_College_Students.pdf (last accessed October 2018)

[9] Retrieved from www.ipu.org/pdf/publications/hr_guide_en.pdf (last accessed October 2018)

[10] Retrieved from www.geocities.ws/eric.engle/GenerationHistory.htm (last accessed October 2018)

[11] Janet Holl Madigan (2004): Being Human, Being Good: The Source and Summit of Human Rights; retrieved from http://drum.lib.umd.edu/bitstream/1903/1753/1/umi-umd-1729.pdf (last accessed October 2018)

that human rights put the individual above the community which goes against the communitarian values.[12]

(46) Indeed, the secular, individualistic and rationalist components of **human rights**, attributable to Western cultural influences …, **are not universal** in the great diversity of cultures, and as such these qualities have the potential to isolate and disengage cultures that are not individualistic, secular or rationalist.[13]

(47) **human rights** are conditioned on capitalism, and if so, they **are not universal**.[14]

The copula sentence is the primordial form of paraphrase. It takes the linguistic form of the lexical item for granted, being a statement on the referent for which the lexical item stands, i.e. the discourse object. This referent is, as I have argued in Teubert (2010), not an object, state, property, process or relationship of the world outside discourse but the object as it has been constructed through negotiations in discourse. We only can talk about human rights because they are (already) objects of our discourse. People without language, or people speaking a language which has no lexical item for this concept or idea would not normally discuss it. The copula sentence does not distinguish between lexical knowledge and encyclopaedic knowledge. As I see it, the meaning of the lexical item and the knowledge we have of the discourse object for which it stands are co-referential. The notion of human rights is what the lexical item *human rights* means.

Paraphrastic knowledge is contained in many other sentence forms. They modify, exemplify, compare or argue something said about a discourse object. Some of these forms are fairly common, and can be retrieved by automated queries. A comprehensive overview of paraphrase patterns is found in Cheung (2009). However, these patterns only point to possible candidates for paraphrastic content. For it is up to the discourse participant, or to the researcher, to decide what counts as such. Paraphrase is not an algorithmic concept. It is always arbitrary. It is possible, for instance, to talk about the issue of universality without ever using the word *universal*. Here are a few examples, found in newspaper archives (search phrase in bold):

[12] "IND AND HUMAN RIGHTS EXAM.doc", retrieved from www.personal.ceu.hu/ (last accessed October 2018)

[13] Retrieved from http://www.e-ir.info/2014/04/25/western-human-rights-in-a-diverse-world-cultural-suppression-or-relativism/ (last accessed October 2018)

[14] Retrieved from https://www.carnegiecouncil.org/publications/archive/dialogue/1_03/articles/514 (last accessed October 2018)

(48) Anyone who is removed from society because of the seriousness of the crime automatically forfeits the rights of that society. But now, **human rights** law turns this upside down by imposing a duty on the state to give people their rights. (Melanie Philips in the *Daily Mail* 2005)

(49) Yes we will scrap the **Human Rights** Act, which has made it incredibly difficult for the government to deport people who they know to be a threat. (David Cameron 2010)

(50) The promotion and protection of **human rights** is at the heart of the UK's foreign policy objectives. I consistently raise human rights violations wherever and whenever they occur. (William Hague in parliament 2012)

(51) The simple truth is that those countries that need **Human Rights** legislation will never accept it while countries such as the UK don't need it. (*Daily Mail* reader comment 2013)

The first two examples deny universality on the level of the individual. The state should not be legally obliged to grant human rights to heinous criminals or immigrants with dangerous thoughts, for such people automatically forfeit them. The third example, on the other hand, declares their universality, everywhere and at all times. The fourth example makes a subtle implicit distinction between human rights as universal law and as legal framework, needed only in regions where the universal law is under assault. Interestingly, a similar controversy was fought out in the House of Commons in respect to the Charter of Fundamental Rights attached to the Treaty of Maastricht (cf. Teubert 2008). According to the Hansard of 5 February 2008, the Conservative MP John Redwood said Britain should reject this charter and remain free to decide on her own legal frame for these rights:

(52) We believe that those rights are best expressed in British law, in the English language and in a way that is answerable to the British people.

The Liberal MP Michael Connarty, however, wants to have the Charter as legally binding, not just in Britain but also in those less fortunate European countries without such a human rights tradition:

(53) The charter will be binding, and I find that attractive for reasons that I will outline. Article 4 of chapter I is entitled, "Prohibition of torture and inhuman or degrading treatment or punishment". Such a prohibition may not exist in some countries that might be considering joining the European Union.

Obviously, the question of universality is only one aspect of human rights. It would be equally interesting to compare the different lists of human (or fundamental) rights, with a view on the arguments for controversial cases. One important aspect is that of power. When the rights of men were first discussed, in the 18th century, it was people deprived of their rights demanding them from the state or the monarch. Today it is governments decrying other governments' human rights violations while denying them to their own people, or it is politicians and the media denying them to those without a voice, immigrants and others considered not to fit in.

While software can be developed to detect typical sentence patterns used for paraphrases, such tools can only throw up candidates which have to be confirmed or rejected by close reading. Close reading is also essential for the kind of paraphrastic content we find outside of common paraphrase patterns. For many corpus linguists, particularly those favouring the corpus-driven approach, selecting what we find through reading, rather than having the computer detect what we are looking for, is anathema. However, in the light of the other factors of subjective interference, such as the compilation of the data and the arbitrary choice between a wide range of statistical tools, we should acknowledge that a strictly scientific methodology cannot be the *ultima ratio* when it comes to meaning. Subjective bias is always counterbalanced by the peer community endorsing or rejecting arbitrary decisions made in the preparation of the textual evidence.

7 Conclusion: Making meaning transparent

The only place to find meaning is discourse. We cannot find it in people's heads, because we cannot look into them. As I have argued in Teubert (2013), all we know, including our first person experiences, is based on what people have told us. It is not a neat formula of an algorithmic nature but what people tell us that satisfies our desire to learn about an object of discourse. The "conceptual or idea" aspect of meaning is constructed by and contained in what people say. In paraphrases, they make clear what it is they talk about. Only rarely they discuss the linguistic form of a lexical item. What they are concerned with are the things people talk about in discourse. If and how the reality of the external world is reflected in discourse we cannot know. Whether there is really a sixth sense does not matter; what matters is what people have to say about it. All relevant human knowledge is knowledge shared in discourse. What we keep to ourselves has no impact. Therefore we must give up the idea that lexical knowledge should be separated from encyclopaedic knowledge. The meaning of a lexical item is co-extensive with the (discourse-internal) knowledge of the discourse object for which it stands.

Corpus linguistics has developed a strong methodology for compiling the textual evidence for the meaning of a given lexical item, for giving it a systematic order and for presenting it to those interested in making sense of it. It uses automatic tools to arrive at an "abstraction at the syntagmatic level". Its contribution to dictionary making has been acknowledged. From the beginning, tools were developed to analyse collocation and pattterns. This has revolutionised our understanding of how language works. But it also meant that linguists and lexicographers continued to exercise their arbitrary authority over the "conceptual or idea" aspect of the meaning of lexical items, thus disenfranchising discourse participants. For it was up to them, to provide dictionary glosses describing the conceptual meaning of lexical items. Therefore we should re-interpret the corpus-driven approach as an admonishment to take note of how language users paraphrase words, expressions and other text segments. Software to identify candidates for paraphrastic content is not out of reach. The same is true for software to discover potential intertextual links showing what each new occurrence adds to the meaning of a lexical item, and where people concur and where they disagree. Once these tasks are accomplished, the work of the linguist is done. For the interpretation of the assembled evidence is not a prerogative of experts, but the democratic right of all discourse participants. Their interpretations will be biased, and thus give room to never-ending discussions in the interpretive community, often leading to new insights. Instead of being told the 'true' meaning of *human rights* by the unassailable authority of the dictionary, it will be the people who construct human rights as they see fit.

Such a scenario is far from utopian. All that needs to be developed is an interface between a vast corpus constantly updating itself, like the web, and the discourse participants. They should be able to decide how much and what kind of evidence they would like to peruse. Some may only want to look at one-sentence definitions of *human rights*, while other may have very specific questions. Interacting with the user, the interface would assemble the evidence, organise it and present it. Rather than attempting to anticipate the user's presumed interests, such an interface would guide them individually to the evidence they are looking for. This is how I see the future of corpus linguistics.

References

Baker, P. (2006). *Using Corpora in Corpora in Discourse Analysis*. London, UK: Continuum.

Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013). *Collocation: Applications and Implications*. Basingstoke, UK: Palgrave.

Cheung, M. L. L. (2009). *Merging Corpus Linguistics and Collaborative Knowledge Construction*. Doctoral dissertation, University of Birmingham, Birmingham, UK. Retrieved from http://etheses.bham.ac.uk/464/1/Cheung09PhD.pdf (last accessed October 2018).

Derrida, J. (1977). *Limited Inc*. Evanston, IL: Northwestern University Press.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448.

Firth, J. R. (1957). Modes of meaning. In J. R. Firth, *Papers in Linguistics 1934 – 1951*. Oxford: Oxford University Press, 190–215.

Fish, S. (1980). *Is there a Text in This Class? The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Galavotti, M. C. (2001). Subjectivism, objectivism and objectivity in Bruno de Finetti's Bayesianism. In D. Corfield & J. Wlliamson (Eds.): *Foundations of Bayesianism*. Dordrecht: Springer, 161–174.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London, UK: Routledge.

Locke, J. (1982). *Second Treatise of Government*. Arlington Heights, IL: Harlan Davidson.

Louw, W. E. (2014). *Collocation as the Determinant of Meaning: From Context of Situation to Corpus-Derived Subtext*. Doctoral dissertation, University of South Brittany, Lorient, France.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Numberwatch (n.d.). Statistical significance. Retrieved from https://web.archive.org/web/20180305164004/http://www.numberwatch.co.uk/significance.htm (last accessed October 2018).

Palmer, H. E. (1933). *Second Interim Report on English Collocations, Submitted to the Tenth Annual Conference of English Teachers under the Auspices of the Institute for Research in English Teaching*. Tokyo, Japan: Institute for Research in English Teaching.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics 10*(1), 1–13.

Teubert, W. (2007). Natural and human rights, work and property in the discourse of Catholic social doctrine. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora* (pp. 89–126). London, UK: Continuum.

Teubert, W. (2008). What is the role of arguments? Fundamental human rights in the age of spin. In E. Weigand (Ed.), *Dialogue and Rhetoric* (pp. 95–118). Amsterdam, Netherlands: John Benjamins.

Teubert, W. (2010). *Meaning, Discourse and Society*. Cambridge: Cambridge University Press.

Teubert, W. (2013). Was there a cat in the garden? Knowledge between discourse and the monadic self. *Language and Dialogue 3*(2), 273–297.

Teubert, W. (2016). Review: Hanks, P. (2013). Lexical Analysis: Norms and Exploitations. Cambridge, MA: MIT Press. (xv + 462 pp.). *International Journal of Corpus Linguistics, 21*(2), 272-283.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam, Netherlands: John Benjamins.

Part II: **Contexts of lexis and grammar**

Gregory Garretson

Family collocation

Exploring relations between lexical families

Abstract: This chapter introduces the concept of 'family collocation', to address a problem with traditional methods of studying collocation: namely, that focusing on the co-occurrence in text of individual forms ignores connections between related forms that are likely to exist in the mental lexicon. For example, the phrases *satisfactory conclusion, satisfactory conclusions, unsatisfactory conclusion*, and *satisfying conclusion* are seen as completely unrelated according to some approaches to collocation, despite the likelihood of speakers seeing connections between them. While some corpus studies have grouped forms by lemma, this method goes one step further and groups related words into 'families'. To demonstrate family collocation, a study is presented of the co-occurrence of 100 high-frequency word families in the British National Corpus. The results suggest that such co-occurrence is very frequent, creating complex sets of relations that should be taken into consideration. Three case studies demonstrate various implications of family collocation for corpus studies.

1 Textual vs. psychological relations

This chapter is concerned with associations between words, both in texts and in the mind. It has two aims: the first is to present some evidence bearing on the long-running discussion about whether collocation is best seen as a textual or as a psychological phenomenon (see e.g. Partington 1998). The second aim is to introduce and illustrate via an exploratory study the concept of 'family collocation', which offers a new way of thinking about lexical relations, in particular about how co-occurrence relations observed in text relate to putative relations in the mental linguistic system.

The following are all excerpts from the British National Corpus (Aston & Burnard 1998, henceforth BNC; three-letter codes denote corpus files):

Gregory Garretson, Uppsala University, gregory.garretson@engelska.uu.se

https://doi.org/10.1515/9783110489071-007

(1) bring the whole matter to a **satisfactory conclusion** as quickly as (AR3)
(2) It was an **unsatisfactory conclusion** to an otherwise profitable visit (GVP)
(3) bringing any of the characters to a **satisfying conclusion** (CBC)
(4) The main and very **unsatisfying conclusion** that I have reached (H7K)
(5) we can then basically come to a **conclusion** which **satisfied** (KN3)
(6) his feeling of intense **satisfaction** at this **conclusion** (HH1)
(7) I reach that **conclusion** with no **satisfaction** (FCH)
(8) anxious for the negotiations [...] to be **concluded satisfactorily** (HH1)

Speakers of English would presumably agree that the word *satisfactory* and the word *conclusion* tend to co-occur. Similar combinations such as *unsatisfactory conclusion* and *satisfying conclusion* are unlikely to seem intuitively very different. Even more distinct pairs such as *conclusion* + *satisfied* and *concluded* + *satisfactorily* seem not unrelated to the original pair of *satisfactory* and *conclusion*. In total, (1)–(8) present no fewer than nine different lemmas representing four different parts of speech[1] – and yet, it is difficult to deny that they have something in common.

This commonality, however, is all but invisible according to most approaches to collocation. This is because collocation is typically seen as the significant co-occurrence of word *forms* in text (where "text" can be spoken or written language). Word forms that are paradigmatically or otherwise related are treated separately, such that similarities in their patterning are only detected if the researcher explicitly looks for them. This view of collocation is in large part the legacy of Sinclair, whose influence on corpus linguistics was fundamental. Sinclair defined collocation as a *textual* relation, stating that "[c]ollocation in its purest sense [...] recognizes only the lexical co-occurrences of words" (Sinclair 1991: 170), and (quite rightly) stressed the importance of allowing the individual profile of each word form to emerge in collocational analysis.

A slightly different approach is to consider the co-occurrence of *lemmas* rather than of *word forms*. One linguist who has worked in both ways is Hoey (2005: 5):

> Collocational analysis can be done on lemmas or words. Renouf (1987), Sinclair (1991), Stubbs (1996), and Tognini-Bonelli (2001) have all argued against conflating items sharing a common lemma (e.g., *political*, *politics*; *break*, *broke*; *onion*, *onions*) on the grounds that each word has its own special collocational behaviour. In Hoey (1991b) and Hoey (1991a) I

[1] These lemmas are as follows: satisfy_VERB, satisfaction_NOUN, satisfactory_ADJ, unsatisfactory_ADJ, satisfactorily_ADV, satisfying_ADJ, unsatisfying_ADJ, conclude_VERB, and conclusion_NOUN. See Section 3 below for the working definition of 'lemma'.

> found it useful to work with lemmas, but for present purposes I concur with these linguists that conflation often disguises collocational patterns.

The choice between working with lemmas and working with word forms may seem fairly trivial. Yet these two approaches are indicative of different views of collocation as a phenomenon. Because what we typically work with are texts, we normally speak of collocation as a *textual* relation. However, the real object of interest for many linguists is not the texts themselves, but rather the linguistic system responsible for their production. As this system exists in the human mind, we might equally say that collocation is a *psychological* relation. In fact, Hoey (2005: 5) explicitly defines collocation as such:

> [O]ur definition of collocation is that it is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution. This definition is intended to pick up on the fact that collocation is a psycholinguistic phenomenon, the evidence for which can be found statistically in computer corpora.

Although this definition is both theoretically and practically justifiable, it is noteworthy that Hoey, while interested in the psychological associations between words, rejects the notion of association between *lemmas*; from this perspective, collocations such as *commits* + *crime* and *committing* + *crime* must necessarily be considered separately.

If what we are truly interested in is lexical associations in the mind, it is important to develop corpus methods that allow us to model these associations. Many attempts have been made in recent years to reconcile the results of corpus linguistic studies of collocation and psycholinguistic studies of word association (e.g. Schmitt, Grandage, & Adolphs 2004; Gries, Hampe, & Schönefeld 2005; Ellis & Simpson-Vlach 2009; Dąbrowska 2014). While in some cases these approaches – corpus-based on one hand and elicitation-based on the other – yield similar results, in many cases they do not (see e.g. Mollin 2009). One possibility, suggested by Nordquist (2004), is that this is because the two methods reflect different types of associations, with elicitation triggering the open-choice principle rather than the idiom principle (Sinclair 1991). Another possibility is that the methods used in corpus linguistics are overly simplistic – while great attention has been paid to the question of which statistical measure to use in collocational studies (e.g. Oakes 1998), somewhat less attention has been paid to the question of what the actual nature of collocation is (though for discussion see e.g. Gries 2008; Garretson 2010).

The modest proposal put forward here is that, without abandoning our investigations of the collocational patterns of individual forms, we broaden our approach. In this way, studies of co-occurrence relations can be expanded to include co-occurrence not only between lemmas, but also between groups of related lemmas such as the ones illustrated in (1)–(8) above, on the grounds that these forms may be similarly associated in the mental linguistic system. This approach is referred to as ‘family collocation’.

2 Co-occurrence across word classes

The idea of grouping inflectionally and derivationally related words, like those in (1)–(8) above, is hardly new. In fact, this was the original view of Firth – who gave us the concept of collocation – and his followers. In 1957, Firth illustrated collocation as follows: “[i]t can be safely stated that part of the ‘meaning’ of cows can be indicated by such collocations as *They are milking the cows*, *Cows give milk*” (in Palmer 1968: 180). Note that in the two examples given, *milk* is in one case a verb and in the other a noun. Similarly, in 1966, Halliday (1966: 150) elaborated on this idea in a discussion of lexical relations that has been cited often over the years:

> The sentence *he put forward a strong argument for it* is acceptable in English; *strong* is a member of that set of items which can be juxtaposed with *argument*, a set which also includes *powerful*. *Strong* does not always stand in this same relation to *powerful*: *he drives a strong car* is, at least relatively, unacceptable, as is *this tea’s too powerful*.

What fewer have remarked upon is that in the same discussion, Halliday (1966: 151) argues that these relations cut across part-of-speech categories:

> *Strong*, *strongly*, *strength*, and *strengthened* can all be regarded for this present purpose as the same item; and *a strong argument*, *he argued strongly*, *the strength of his argument* and *his argument was strengthened* all as instances of one and the same syntagmatic relation.

We may say, therefore, that for Firth and Halliday, the idea that collocation might cut across grammatical categories was relatively unproblematic. Since then, this perspective has gradually lost ground among those working with collocation, though there has been relatively little discussion of the theoretical arguments for this shift in perspective.

Of course, some more recent research, such as Nesselhauf (2005: 17), does take the approach of allowing collocation to apply to lemmas, but stops short of allowing it to cross part-of-speech categories:

> Combinations such as *strong argument* and *strong arguments* are therefore generally assumed to be instantiations of the same collocation, but *the strength of the argument* is assumed to be an instantiation of a different one.

At the same time, there are linguists who have focused explicitly on such category-crossing relations. For example, Kövecses (1986: 133) refers in his work on conceptual metaphors to “the collocability of concepts”, stating that “[i]t is clear that it is possible to formulate a rule for the collocations of these items which would be on a more general level than the words that form a part of the collocations”. While Kövecses’s approach is not typical of the methodologies of corpus linguistics, considering such views might enrich our understanding of the relations hidden behind the surface patterns in corpus data.

Within the corpus-linguistic tradition, there are also linguists who have noted this phenomenon but dedicated little attention to it. For example, Stubbs (2002: 30) makes the following observation:

> These syntagmatic co-occurrence relations often cross-cut the way in which dictionaries have traditionally represented head-words. Sometimes different forms of a lemma behave differently […], but sometimes forms which are usually regarded as separate lemmas behave similarly. One such case is the collocational relation between the lemmas ARGUE and HEAT. One finds
>
> > argue heatedly; heated argument; in the heat of the argument
>
> These phrases cross-cut the traditional parts of speech […]. In this case, the collocation is between semantic units, irrespective of grammatical category; but there is still a restriction on word-form, since the form *heat* has to occur: *heated argument*, but not **hot argument*.

While Stubbs’s (2002) point is well made, his postulated restriction on the relation appears not to hold. The BNC provides examples such as the following:

(9) Other reports suggest, plausibly, that the hottest argument now is about money. (ABG)

(10) Much of the history of music consists of often hotly argued opinions about what music actually consists of. (J1A)

This suggests that, if the relation represented in these examples is indeed the same as that discussed by Stubbs (2002), such relations can in fact obtain across different lemmas and different roots. But how commonly do such crossings occur? Are they important enough that linguists should pay attention to them? The goal of this chapter is to argue that such category-crossing relations are in fact of great importance, and that – especially if we wish our methods of collocational analysis to have a bearing on linguistic associations in the mind – it would be wise to devote more attention to these relations.

3 Family collocation

Let us hypothesise that if there is a mental association between word pairs such as *strong* and *argument*, there is likely also to be a mental association between related word pairs such as *strength* and *argument* or *strong* and *argumentation*. The task is then to look for textual evidence of such relations. Below, a methodology for finding sets of relations between families of words will be presented, followed by some highly suggestive results.

First, however, it is necessary to define some important terms. It is difficult to describe the super-categorical syntagmatic relations discussed here, since relations of inflection, derivation, and even synonymy may be involved. The term adopted here is ‘lexical family’, or simply ‘family’. For present purposes, we may define a family as a set of lemmas (subsuming their various inflected forms) that are putatively psychologically related in meaning. These lemmas are likely to be related derivationally, though exactly how close two forms must be in terms of etymology and phonetic form to be included in the same family is a question that will require further study. It is possible that semantic relatedness may play as important a role as etymological relatedness in forming mental associations (see e.g. the UCREL Semantic Analysis System described in Rayson et al. 2004). However, for the purposes of the present exploratory study, we will restrict families to sets of (transparently) etymologically related words. Two examples, showing surface forms, are given in (11) and (12):

(11) *know*, *knows*, *knew*, *known*, *knowing*, *knowingly*, *knowledge*, *knowledgeable*, *unknown*, *unknowing*, *unknowingly*, *unknowable*, etc.

(12) *live*, *lives*, *lived*, *living*, *life*, *lifeless*, *lifelike*, *lively*, *liveliness*, *alive*, *enliven*, etc.

In fact, it is convenient to shorten the lists by including only the *lemmas* and allowing software to look for all inflected forms. This will be done in the families presented below.

While the ‘lexical families’ used here are similar to the ‘word families’ described in Bauer and Nation (1993) and used in subsequent work including Coxhead’s (2000) Academic Word List, there are some differences to note. First, the families created for the present study involve a slightly different set of affixes (see below) than Bauer and Nation's (1993) complete set of level 1–6 affixes (and include some affixes not contained in Bauer and Nation's (1993) set, such as *over-*). Second, as described below, lower-frequency lemmas, regardless of form, were excluded from the families in this study for practical purposes. In general, while Bauer and Nation's (1993) procedure for creating word families serves as a useful model, its primary aim is to inform studies of vocabulary development, and not to model associations in the mental lexicon of adult native speakers. For this reason, it can be expected that as evidence for family collocation accumulates, the definition of lexical family used in collocational studies will gradually diverge from Bauer and Nation's.

It is also prudent to define the term 'lemma'. While this term has been used in various ways, here it is used to refer to the pairing of a base form and a part of speech, notated below like so: *live_VERB*. This is considered to subsume the forms *live*, *lives*, *lived*, and *living*, when these are verbal forms. By contrast, the noun *lives* (note the difference in pronunciation) is a form of the lemma *life_NOUN*, while the noun *living* is a form of the lemma *living_NOUN*.[2]

This definition of lemma has two particular limitations that should be noted. First, we may not assume that a given lemma is restricted to only one sense. That is, a lemma may be polysemous: *right_ADJ* may mean either "not left" or "not wrong". The second limitation of this definition is that its usefulness depends crucially on the accuracy of part-of-speech tagging software. The data presented here comes from the BNC, with tags already applied by version 4 of the CLAWS tagger (Garside & Smith 1997). While this tagger does feature a very high accuracy rate, it is not error-free, so some words will have been tagged with the wrong lemma.

If a family is a group of related lemmas, then the analysis of family collocation is the analysis of all collocation between *any two word forms representing two particular families*. The results are grouped both by lemma and by family. For example, given the two families illustrated in (11) and (12) above, we might find that in some corpus they collocate in the following ways:[3]

[2] Words with the code NOUN originally had the code SUBST (for "substantive") in the CLAWS parser tags (Garside & Smith 1997); this has been changed here for reasons of transparency.

[3] Note that the lemmas in the pairs are not presented in the linear order found in the corpus, but are rather arranged such that the members of one family are always on the left and the members of the other are on the right.

(13) *know_VERB + life_NOUN*
know_VERB + live_VERB
know_VERB + alive_ADJ
know_VERB + living_ADJ
know_VERB + living_NOUN
know_VERB + live_ADJ
knowledge_NOUN + life_NOUN
knowledge_NOUN + live_VERB
unknown_ADJ + life_NOUN
known_ADJ + life_NOUN
known_ADJ + live_VERB
etc.

In order to get a sense of the prevalence of family collocation, an investigation was carried out in which the BNC was searched for co-occurrences of the members of 100 high-frequency families. The next section presents the methodology used, and the subsequent section presents the results obtained.

4 Constructing and relating families

The study reported here examines the co-occurrence relations among 100 lexical families in the BNC. As the study is exploratory in nature, the results are more qualitative than quantitative. This section presents the methodology used, which comprises two main steps: constructing the families and looking for co-occurrence relations between them.

4.1 Constructing the families

The XML version of the BNC (Burnard 2007) was chosen as the material for the study, for three reasons. First, it is a comparatively clean and well-described corpus. Second, at 100 million words, it is certain to yield sufficient data. Third, thanks to the CLAWS tagger (Garside & Smith 1997), all of the words in the corpus are lemmatised; there is a headword (“hw”) attribute for each one, as shown in (14):

(14) <w c5="PNP" hw="they" pos="PRON">They </w><w c5="VBB" hw="be" pos="VERB">are </w><w c5="VVN" hw="force" pos="VERB">forced </w> (EVA)

This means that instead of searching for individual inflected forms and then grouping them by lemma, it was possible to search by lemma. However, because the definition of lemma used in this study involves both a base form and a part of speech (see above), when constructing word lists, the words' "hw" and "pos" attributes were combined, yielding, for (14), *they_PRON*, *be_VERB*, and *force_VERB*. Note that this last lemma differs from the lemma *force_NOUN*; using the POS information provided by the tagger greatly reduced the problem of homographs, as discussed below.

The first step in constructing the families was to create a word frequency list for the corpus – or in this case, a lemma frequency list. This included over 300,000 lemmas occurring more than once. The top five hundred lemmas were examined, and a set of potentially interesting ones was selected (categories such as prepositions and conjunctions were excluded). A Perl program was written that took as input a list of the selected base forms (or more precisely, hand-crafted search terms based on these) and searched the lemma frequency list for potentially related lemmas, based on spelling similarities, and allowing any of the derivational prefixes shown in (15). The resulting sets of words were manually edited.

(15) *anti-*, *co-*, *com-*, *con-*, *contra-*, *de-*, *dis-*, *en-*, *ex-*, *mis-*, *im-*, *in-*, *inter-*, *over-*, *pre-*, *pro-*, *re-*, *sub-*, *super-*, *trans-*, *un-*, *under-*

The following restrictions were applied to the families: (a) only nouns, verbs, adjectives and adverbs were included; (b) the members of a family had to be transparently related both etymologically and in meaning; (c) only the derivational prefixes listed in (15) were allowed; and (d) no compounds were included.

Naturally, the criterion of lemmas being *transparently* related was somewhat problematic and relied upon the researcher's judgment. When in doubt, the policy was to be conservative and exclude candidates. Some words were excluded on the basis of the etymological relation being not fully clear – for example, the family including *state_NOUN* does not include *stately_ADJ* or *statistic_NOUN*. Others were excluded on the basis of the meaning relation not being fully clear – for example, the family including *real_ADJ* does not include *realise_VERB*, as the most common sense of this verb ("become aware of") is not obviously related in meaning to *real_ADJ*.

One obstacle to the creation of families is homography. For example, *close_VERB* and *close_ADJ*, though pronounced differently, are spelled identically. This is where the use of POS tags was extremely helpful. Although there are certain lemmas that are simply polysemous, such as *close_VERB* (which can mean either "shut" or "draw nearer"), it is possible to separate the two different semantic fields to which *close_VERB* belongs to a large degree: compare *close_ADJ*, *closely_ADV*, *closeness_NOUN* on one hand with *close_NOUN*, *closed_ADJ*, *closure_NOUN* on the other. These lemmas were assigned to separate families (one of which was subsequently excluded from the study).

The exclusion of compounds was seen as a necessary step for this exploratory study, though of course how to treat them is an important question for future work. For example, the words *lightbulb* and *flashlight* are both transparently related to *light_NOUN*, though each might belong to another family as well. Whether to allow compounds to belong to multiple families is a theoretical and practical question that merits discussion.

Tab. 1: Names of the 100 families included in the study, sorted alphabetically.

ABLE
BELIEVE
CALL
CARE
CENTRE
CLEAR
CLOSE
COMPARE
CRIME
DEPARTMENT
DIE
DIFFERENT
EASY
ECONOMY
ELECTION
EMPLOY
END
EXAMPLE
EXPERIENCE
FACE
FAIR
FEEL
FINANCIAL
FIND
FOLLOW
FUND
GOOD
GOVERN
GREAT
GROUP
HAND
HARD
HELP
HIGH
HISTORY
HOLD
HOUSE
HUMAN
IMPORTANT
INFORM
JUDGE
JUSTICE
KNOW
LAST
LEAD
LEGAL
LIGHT
LIMIT
LIVE
LOCAL
LONG
MANAGE
MARRY
MEAN
MIND
NATION
NATURE
NEED
NEW
NORMAL
NUMBER
OFFICE
OLD
PART
PAY
PERIOD
PLACE
PLAN
PLEASE
POINT
POOR
PROBABLE
PROBLEM
PROJECT
PROPOSE
PROVIDE
REAL
REASON
RIGHT
RUN
SAVE
SERVE
SHORT
SKILL
SOCIAL
SPEND
START
STATEMENT
SUPPORT
SURPRISE
SYSTEM
THINK
TIME
USE
WAIT
WALK
WANT
WEEK
WORK
WRITE

Once a large number of families had been created, 100 were selected for the study. Because the goal was to examine frequently occurring and internally diverse families, a number of criteria were applied. Any family that met the following criteria was included: (a) after all lemmas occurring in the corpus fewer than 40 times were pruned from the family, it had to have at least four lemmas left; (b) at least one of these had to occur at least 4000 times in the corpus; and (c) the family had to include at least two lemmas (with different parts of speech) occurring at least 400 times each. While these restrictions are admittedly arbitrary, they were found to result in an appropriate number of high-frequency families. Each of the 100 families chosen for the study was given a simple name (written in capitals); these are listed in Table 1.

The families varied considerably in size and internal diversity. With all members occurring fewer than 40 times in the BNC removed (see above), the families ranged in size from four (the floor imposed) to 27 members, with a median size of 7. All in all, there were 858 lemmas across the 100 families, corresponding to 1520 different surface forms in the corpus.

4.2 Looking for relations between families

Once the set of families to be used in the study had been determined, the next step was to look for co-occurrences. As an illustration of family co-occurrence, consider the two families in (16), for each of which only three members are shown:

(16) SERVICE: service_NOUN, serve_VERB, service_VERB
PROVIDE: provide_VERB, provision_NOUN, provider_NOUN

Co-occurrence of these lemmas in SERVICE and PROVIDE could be realised by any (and possibly all) of the combinations shown in (17). In the interest of clarity, the members of a family are always presented on the same side (left or right):

(17) service_NOUN + provide_VERB
service_NOUN + provision_NOUN
service_NOUN + provider_NOUN
serve_VERB + provide_VERB
serve_VERB + provision_NOUN
serve_VERB + provider_NOUN
service_VERB + provide_VERB
service_VERB + provision_NOUN
service_VERB + provider_NOUN

In the text itself, these are realised in various ways:

(18) to provide a service (KCF)
the services that are being provided (J3R)
the provision of services (J9B)
employers and service providers (AHX)
the present level of service provision (JNB)
as a provider of services (HXT)
all serve to provide a sound backdrop (FT5)
served to provide votes (CCR)
provided ready to serve by the local supermarket (CDN)
etc.

Note that not all of these examples necessarily correspond to the same underlying semantic relation between SERVE and PROVIDE; this, however, is true of collocation in general.

A common practice when studying collocation (see e.g. Sinclair 2004), is to use a span of four words to the left and four words to the right of the node (the node being the word which serves as the starting point of the search). This is not the only possible approach, however; some researchers have used the far more stringent requirement that the two words be in a direct grammatical relationship (see e.g. the discussion in Garretson 2010). Whether this is necessary has been debated at least since Halliday's aforementioned text from 1966 (151–152):

> As far as the collocational relation of *strong* and *argue* is concerned, it is not merely the particular relation into which these items enter that is irrelevant; it may also be irrelevant whether they enter into any grammatical relation with each other or not. [...] Clearly there are limits of relevance to be set to a collocational span of this kind; but the question here is whether such limits can usefully be defined grammatically, and it is not easy to see how they can.

Whether the links formed between word forms (or lemmas) in the mental lexicon are in fact sensitive to direct grammatical relations alone, or show more flexibility, is an important, though difficult, question to investigate. Until a satisfactory answer has been arrived at, it seems prudent to keep both possibilities open. In the present study, the practice followed was to use the traditional nine-word window, with the restriction that only words in the same sentence were considered.

(19)

43470	SOCIAL		
4156	WORK		
		2004	social_ADJ + worker_NOUN
		1594	social_ADJ + work_NOUN
		165	social_ADJ + work_VERB
		29	social_ADJ + works_NOUN
		24	social_ADJ + working_NOUN
		23	social_ADJ + working_ADJ
		2	social_ADJ + overworked_ADJ
		100	society_NOUN + work_NOUN
		73	society_NOUN + work_VERB
		28	society_NOUN + worker_NOUN
		22	society_NOUN + working_ADJ
		14	society_NOUN + works_NOUN
		13	society_NOUN + working_NOUN
		18	socially_ADV + work_NOUN
		4	socially_ADV + work_VERB
		3	socially_ADV + worker_NOUN
		1	socially_ADV + working_ADJ
		2	socialization_NOUN + work_NOUN
		1	socialization_NOUN + worker_NOUN
		2	societal_ADJ + work_VERB
		1	societal_ADJ + work_NOUN
		1	anti-social_ADJ + work_NOUN
		1	sociable_ADJ + work_NOUN
		7	socialise_VERB + work_NOUN
		etc.	

In order to find evidence of the co-occurrence of families, another Perl program was written that searched the BNC for the co-occurrence of any two words on a given list. The input consisted of the 858 lemmas of the 100 families, grouped by family. For each lemma, the program searched the corpus for every word tagged as representing that lemma. Each time such a word was found, the eight closest words (or fewer, if the node was near a sentence boundary) were checked to determine whether any of them were also on the list. If so, this was recorded. Once this process was complete, the program output a list of co-occurrences like that shown (truncated slightly) in (19) for the families SOCIAL and WORK.

There are several things to notice about this list. First, no statistical tests have been performed as yet; the list gives raw counts of co-occurrences. Second, the words shown are lemmas, not the actual surface forms that occurred in the corpus. Third, the linear order in which the lemmas are shown on each line is not necessarily that found in the corpus. Fourth, the vertical ordering of the pairs is as follows: first comes the most frequent lemma from the 'node' family (*social_ADJ*), together with each of the forms from the 'collocate' family with which it co-occurs (*worker_NOUN*, *work_NOUN*, etc.), in descending order of frequency of co-occurrence. Then follows the second-most-frequent lemma from the node family, with the lemmas it co-occurs with, and so on.[4] The next section presents the most interesting aspects of these results, as well as some case studies of family collocation.

5 Family collocation illustrated

To give a sense of how prevalent the phenomenon of family collocation is, this section will present some general results before turning to three case studies which will be explored in greater detail.

5.1 General results

The 100 lexical families studied, comprising 858 lemmas corresponding to over 1500 word forms, accounted for approximately 5,700,000 of the 98,400,000 words in the corpus – that is, roughly 6% of the corpus. This relatively high proportion is not surprising, as these families were selected to include a large number of frequent words.

The program measured not only the number of co-occurrences of lemmas from different families, but also how many *opportunities* there were for such lemmas to co-occur. That is, for each and every token of some word on the list, it counted the nouns, verbs, adjectives and adverbs within four words to the left and right (stopping at sentence boundaries). This information, recorded for each

[4] The numbers at the top tell us that the family SOCIAL co-occurred with all 100 families 43,470 times, and 4,156 of these instances (roughly one-tenth) were co-occurrences with the family WORK.

family, represents the total number of opportunities another family had to co-occur with it. The total number of such ‘slots’ was approximately 23 million, roughly one quarter of the corpus. This means that on average, every node word found was surrounded by four content words with the potential for affecting family collocation.

Family collocation was found to cover a considerable portion of the corpus. Each of the 100 families co-occurred with every other family (plus itself), indicating that these words form a dense network. Of the 23 million ‘neighbour slots’, as many as 10% were filled by some representative of one of these 100 families.

Even more interesting, and more important for current purposes, is the variety of ways in which families were found to co-occur. Example (20) shows the interaction of two relatively small families with only four members each: IMPORTANT and PROVIDE.

(20)

19921	IMPORTANT		
392	PROVIDE		
		255	important_ADJ + provide_VERB
		53	important_ADJ + provision_NOUN
		5	important_ADJ + provider_NOUN
		2	important_ADJ + provisional_ADJ
		41	importance_NOUN + provide_VERB
		13	importance_NOUN + provision_NOUN
		2	importance_NOUN + provider_NOUN
		2	importance_NOUN + provisional_ADJ
		16	importantly_ADV + provide_VERB
		2	importantly_ADV + provision_NOUN
		1	unimportant_ADJ + provide_VERB

Since each of these families contains four members, there are sixteen possible combinations (4 × 4), eleven of which were found to occur in the corpus. Using such numbers, we can calculate a simple measure of the diversity of interaction of any two families that takes into account the fact that families differ in size. Let us define the ‘diversity score’ as the proportion of the possible lemma combinations for two families that actually occurs in the corpus. For example, the diversity score for IMPORTANT and PROVIDE is 11 ÷ (4 × 4) = 0.69, meaning that 69% of the possible combinations (the eleven shown above) were found. Because this metric privileges pairs of families with fewer members, it will be reported together with the sizes of the families.

Two points bear repeating here. The first is that no claim is made that all of the co-occurrences of these lemmas in the text represent the same semantic relation. The other is that these are raw co-occurrence numbers; none of these relations have as yet been submitted to a test of statistical significance. The idea is that while the combination of forms *important + provide* may yield a higher MI or log-likelihood score than *importance + providing* (this last form being subsumed under *provide_VERB*), the repeated occurrence of each of these pairs is likely to strengthen the mental association represented by the other. Thus, any statistical tests should be performed on the *sum* of all combinations encountered.[5]

If we calculate a diversity score for all 10,000 possible combinations of the 100 families, we find that these range from 1.00 at the top (meaning that *all* possible combinations occur) to 0.009 at the bottom. However, of the 10,000 combinations, only one (LEAD + LEAD) is over 0.95, and only one (LEGAL + PLEASE) is under 0.01. Most pairs have a diversity score of between 0.10 and 0.30, meaning that 10%–30% of all possible combinations of lemmas occurred. Note that this says nothing about *how often* the families co-occur; it merely says something about *how many different ways* they co-occur.

The average diversity score for the 10,000 combinations is 0.212. This means that overall, one-fifth of the possible combinations of lemmas occur. A typical case is shown in (21). The family ABLE contains 13 different lemmas, and the family MARRY contains 8, yielding 104 possible combinations. Of these, 21 different combinations were in fact found in the corpus, giving this combination a diversity score of 0.20.

If the average diversity score is 0.212, we may ask how many pairs of lemmas that typically corresponds to. The average number of lemma-lemma co-occurrence pairs between any two of the 100 families is 13, though the results for the 10,000 pairs range from as few as 1 to as many as 101 unique pairs of lemmas. Again, it is worth bearing in mind that the actual surface forms show a much greater diversity, as will be illustrated below.

[5] While this line of reasoning could be seen as arguing against pruning low-frequency lemmas from the families (see above), choosing a floor was judged to be a reasonable way of focusing on lemmas likely to make a difference.

(21)

27594	ABLE		
72	MARRY		
		22	able_ADJ + marry_VERB
		6	able_ADJ + married_ADJ
		5	able_ADJ + marriage_NOUN
		1	able_ADJ + unmarried_ADJ
		4	ability_NOUN + marriage_NOUN
		2	ability_NOUN + marry_VERB
		4	enable_VERB + marry_VERB
		3	enable_VERB + marriage_NOUN
		2	enable_VERB + married_ADJ
		1	enable_VERB + marital_ADJ
		6	unable_ADJ + marriage_NOUN
		3	unable_ADJ + marry_VERB
		2	unable_ADJ + married_ADJ
		1	unable_ADJ + remarry_VERB
		1	disabled_ADJ + married_ADJ
		1	disabled_ADJ + marry_VERB
		3	disability_NOUN + marital_ADJ
		2	disability_NOUN + married_ADJ
		1	disability_NOUN + marriage_NOUN
		1	inability_NOUN + married_ADJ
		1	disablement_NOUN + marital_ADJ

It seems clear, then, that given 100 high-frequency families in a fairly large corpus, these families will interact in a wide diversity of ways. We may hypothesise that this frequent co-occurrence of closely related lexical items in language will have an effect on the mental associations formed. The next sections present a closer look at some examples and then consider the implications of lexical families for studies of collocation.

5.2 Case study 1: LIVE and DIE

It should come as no surprise that the antonyms *life* and *death* collocate in English text. Indeed, some scholars claim that co-occurrence in text is crucial for the development of antonymic relations (e.g. Justeson & Katz 1992). And in fact, the words *life* and *death* co-occur in the BNC over 500 times. What is perhaps more surprising is the vast complex of co-occurring words related to *life* and *death*. Example (22) shows the 35 different ways in which the family LIVE co-occurs with the family DIE in the BNC. These 35 pairs of lemmas, representing a diversity score of 0.45 (LIVE having 13 members and DIE 6), co-occur in the corpus 1406 times.

(22)

48893	LIVE		
1406	DIE		
		587	life_NOUN + death_NOUN
		97	life_NOUN + die_VERB
		44	life_NOUN + dead_ADJ
		7	life_NOUN + dying_ADJ
		6	life_NOUN + deadly_ADJ
		1	life_NOUN + deathly_ADJ
		185	live_VERB + die_VERB
		54	live_VERB + dead_ADJ
		42	live_VERB + death_NOUN
		4	live_VERB + dying_ADJ
		54	living_NOUN + dead_ADJ
		7	living_NOUN + dying_ADJ
		6	living_NOUN + death_NOUN
		5	living_NOUN + die_VERB
		148	alive_ADJ + dead_ADJ
		19	alive_ADJ + die_VERB
		8	alive_ADJ + death_NOUN
		1	alive_ADJ + deadly_ADJ
		1	alive_ADJ + dying_ADJ
		33	living_ADJ + death_NOUN
		30	living_ADJ + dead_ADJ
		8	living_ADJ + die_VERB

16	live_ADJ + dead_ADJ
16	live_ADJ + death_NOUN
11	live_ADJ + die_VERB
2	live_ADJ + deadly_ADJ
2	lively_ADJ + death_NOUN
1	lively_ADJ + dead_ADJ
2	lifeless_ADJ + dead_ADJ
2	lifeless_ADJ + death_NOUN
1	lifeless_ADJ + die_VERB
2	relive_VERB + death_NOUN
2	relive_VERB + dying_ADJ
1	relive_VERB + dead_ADJ
1	enliven_VERB + dead_ADJ

Predictably, the most frequent combinations are *life* and *death*, *live* and *die*, and *alive* and *dead*. However, there are many others; we also find *life* and *die*, *live* and *death*, *living* and *death*, etc. But again, these are lemmas, which subsume several paradigmatically related forms such as plurals and past tense forms. In short, these two families co-occur in a rich variety of different patterns, as exemplified in (23)–(32).

(23) Envy-management, in short, can be life or death for a society. (A69)
(24) And she had chosen to die as she had lived. (FNT)
(25) ‘Much rather be alive than dead.’ (B7J)
(26) ‘You saved my life when I was dying on the hillside.’ (HGS)
(27) Living with the dead is my life. (CA3)
(28) Is it possible to put life into dead things? (H8G)
(29) [...] France, where he lived until his death in l655. (ABM)
(30) Life without hope is a living death. (B21)
(31) Here the living and the dead were in harmony. (ASE)
(32) Finally assured that he had been truly alive, he died. (GVL)

A list such as this makes possible an interesting mental exercise – namely, to ask oneself, reading the examples, whether they seem to embody one association or ten different ones. According to the typical approach to collocation, these examples represent ten out of the several dozen different (and a priori unrelated) associations between the LIVE and DIE families, some of which will turn out to be statistically significant, and others not. The suggestion offered here is that there is a level at which these myriad instances of lexical co-occurrence serve to reinforce each other and build a more general semantic association between the family LIVE and the family DIE.

5.3 Case study 2: WORK and EMPLOY

Two other families that are clearly related semantically are WORK and EMPLOY. But unlike ‘classic’ examples of collocation such as *cause* + *problem* and *commit* + *crime*, there is no one pair of words from these two families that immediately leaps out as the most typical; each family features both frequent nouns and frequent verbs. Accordingly, there are several highly frequent combinations, such as *employ* + *worker*, *employee* + *work*, and *employer* + *worker*. In fact, these two families co-occur in the BNC 1729 times in no fewer than 46 different ways (yielding a diversity score of 0.39, WORK having 13 members and EMPLOY 9). Example (33) shows the 24 most common of these pairs.

(33)

74429	WORK		
1729	EMPLOY		
		98	work_NOUN + employee_NOUN
		98	work_NOUN + employment_NOUN
		76	work_NOUN + unemployed_ADJ
		74	work_NOUN + employer_NOUN
		64	work_NOUN + employ_VERB
		29	work_NOUN + unemployment_NOUN
		145	work_VERB + employee_NOUN
		97	work_VERB + employer_NOUN
		49	work_VERB + employ_VERB
		40	work_VERB + employment_NOUN
		32	work_VERB + unemployed_ADJ
		14	work_VERB + unemployment_NOUN
		218	worker_NOUN + employ_VERB
		180	worker_NOUN + employer_NOUN
		121	worker_NOUN + employment_NOUN
		117	worker_NOUN + unemployed_ADJ
		52	worker_NOUN + employee_NOUN
		38	worker_NOUN + unemployment_NOUN

18	worker_NOUN + employed_ADJ
11	works_NOUN + employ_VERB
14	working_ADJ + unemployed_ADJ
11	working_ADJ + employ_VERB
28	working_NOUN + employee_NOUN
21	working_NOUN + employment_NOUN

These pairs of lemmas illustrate an important point about corpus data and the study of collocation. Some of these base forms, such as *work*, can belong to two different parts of speech – in this case either verb or noun. Any study of collocation that does not make use of part-of-speech tagging (or some other more sophisticated markup) will necessarily conflate the associations of the verb *work* and those of the noun *work*, while in all likelihood keeping separate the associations of the verb forms *work* and *works*. This would seem to be in conflict with the stated goal of examining the associations of each form individually. If it is true that "each word has its own special collocational behaviour" (Hoey 2005: 5; see above), then surely the verb *work* and the noun *work* both have their own special collocational patterns, and lumping these together is less principled than it would be to lump together the verb forms *work* and *works*.[6] This suggests that, even if one does not deal with lemmas *per se*, using part-of-speech information may have a crucial effect on the results of collocational studies.

We may repeat with examples (34)–(43) the mental exercise mentioned earlier: does there appear to be a level at which all of these forms participate in one global association, or are these clearly different associations?

(34) The probability of any worker being employed (H9A)
(35) problems that may arise between the disabled worker and his employer (FS6)
(36) policy demands that an employee be free to work for whom he chooses (J7B)
(37) experience that enables the worker to enter mainstream employment (B0N)
(38) but this takes time and may be costly for the unemployed worker (BNW)
(39) admitted taking heroin with one employee after work (K5M)

[6] Similarly, a study of the form *works* is likely to conflate the verb *works* and the noun *works*.

(40) those who are frustrated in their search for work and employment (GX0)
(41) were more likely to work for an employer who (FAF)
(42) leaving the potential work force not only unemployed but unemployable (J9L)
(43) when he depends upon his employer for his future work (JNP)

The ten examples above include eight different surface forms and three different parts of speech. A full illustration of all 46 lemma combinations for WORK and EMPLOY would feature over 25 distinct words, all tied to each other in a complex lexical network.

5.4 Case study 3: NATION and LEAD

The third and final case study concerns the families NATION and LEAD. While the family LEAD contains a modest five members, NATION contains no fewer than twenty-five different lemmas; these families co-occur 837 times in 48 different combinations (yielding a diversity score of 0.38, similar to that for WORK and EMPLOY). The 18 most frequent of these are shown in (44).

(44)

33696	NATION		
837	LEAD		
		207	national_ADJ + leader_NOUN
		134	national_ADJ + lead_VERB
		59	national_ADJ + leadership_NOUN
		30	national_ADJ + leading_ADJ
		11	national_ADJ + lead_NOUN
		58	international_ADJ + lead_VERB
		38	international_ADJ + leading_ADJ
		27	international_ADJ + leader_NOUN
		12	international_ADJ + lead_NOUN
		10	international_ADJ + leadership_NOUN
		36	nation_NOUN + lead_VERB
		35	nation_NOUN + leader_NOUN
		31	nation_NOUN + leading_ADJ
		14	nationalist_NOUN + leader_NOUN
		12	nationalist_NOUN + lead_VERB

33	nationalist_ADJ + leader_NOUN
10	nationalist_ADJ + lead_VERB
13	nationally_ADV + lead_VERB

An interesting characteristic of this pair of families is the wide variety of combinations it features – from *national leader* to *leading international expert*, to *leading industrial nation*, and even *nationalist leader*. The ten most frequent combinations are illustrated in (45)–(54).

(45) Mr Kinnock will make a good, or even an adequate, national leader (AHN)
(46) a demand to arm and lead a national revolution against the French (FB7)
(47) a call for his return to national leadership (CR9)
(48) that Britain would lead the international commitment vigorously (K57)
(49) regarded as a leading international expert on Roman mosaics. (AKS)
(50) First class sport in County Durham is leading the nation (K52)
(51) The leaders of our nation do however have a charge to stimulate (G0C)
(52) The nationalist leader often had to invent a unifying culture (ANT)
(53) made the USA the leading industrial nation in the world (EWG)
(54) most of the institutions [...] are national or leading research bodies (J0V)

While there is an undeniable semantic consistency between *lead the nation*, *leaders of the nation*, and *national leaders*, the overall diversity of the examples here points to a tension: the more liberal and inclusive we are in defining families, the more variety we can expect in the semantic relations between them. It may therefore be desirable, when studying family collocation, to place limits on families, either in terms of size or in terms of derivational complexity, or both.

6 Implications of family collocation

Having seen that lexical families participate in dense and varied networks of co-occurrence, we may ask what implications this has for studies of collocation. This section will make a first attempt to answer this question, although a thorough investigation of family collocation is beyond the scope of this exploratory study. After briefly considering some of the theoretical implications of family collocation, we will look at the practical question of how family collocation can produce results different from those of traditional approaches to collocation.

Formulaic language, which can be seen as a superordinate category including collocation, phraseology, and related phenomena, has recently been the subject of intensive investigation in various fields – see, for example, Granger and Meunier (2008), Wulff and Titone (2014), and especially the review of the literature in Wray (2012). The various approaches to formulaic language employ highly varied methodologies, ranging from corpus studies to reaction-time experiments to electroencephalography; and indeed, Wray (2012) raises the possibility that these different strands of research are in fact examining different phenomena. Yet a striking commonality among the majority of these studies, whether they be corpus-based or psycholinguistic, is a tendency to work with specific word forms. Corpus approaches, certainly, vary between methods which study highly fixed strings, such as lexical bundles (e.g. Biber et al. 1999; Biber, Conrad, & Cortes 2004), and methods which employ constructs allowing for more variation, such as concgrams (e.g. Cheng, Greaves, & Warren 2006; Cheng et al. 2009). Nevertheless, while the analysis of ROLE/PLAY presented in Cheng et al. (2009), for example, does indeed work with lemmas (*plays/played/playing* + *role/roles*), it excludes cases in which the syntactic categories are changed (e.g. *role play*). Furthermore, although it might seem an obvious question, Cheng et al. (2009) do not discuss whether *play a role* is synonymous with *play a part*, such that the two expressions should be studied together. In all fairness, however, consideration of such lexical alternation in corpus studies is rare.

The point to be made is that our current methods are fairly restrictive in terms of what they look for, and thus in terms of what they can find. The proposal put forward here is that loosening these restrictions may allow us to find evidence of broader associations in language than we have been able to identify thus far. Family collocation represents one way of expanding the domain of study, by going beyond single word forms to include sets of related words.

Now let us turn to the practical question of how the results of a study of family collocation might differ from those of typical collocational studies. To do this, we will consider the three case studies presented above from a statistical perspective. There are many statistical tests used for measuring collocation; for this comparison, log-likelihood (Dunning 1993) was selected because it is in widespread use, has been shown to perform well, and is less sensitive to variation in frequency than mutual information, or MI (Oakes 1998: 174). That said, different tests will give different results, so more thorough testing will be necessary in future studies.

Table 2 shows data for individual forms representing the families LIVE and DIE – specifically, the forms found in examples (23)–(32). In all cases, the pairs now show surface forms; even though this will lead to some conflation of different parts of speech (see the discussion above), this approach is standard enough that it is adopted here. For each pair of forms, the table presents the frequency of each form in the BNC, the number of times they co-occur (within four words, not crossing sentence boundaries), and the resulting log-likelihood score.[7]

Tab. 2: Frequencies and log-likelihood scores for pairs of surface forms representing the families LIVE and DIE.

A + B	Freq. of A	Freq. of B	Co-occurrence of A + B	Log-likelihood score
life + death	54907	19856	531	1263.44
alive + dead	4033	11857	140	811.22
living + dead	15553	11857	105	270.19
live + die	16821	5305	77	257.28
live + dead	16821	11857	37	29.74
living + death	15553	19856	48	27.79
alive + die	4033	5305	3	1.60
life + die	54907	5305	20	0.16
life + dead	54907	11857	37	−0.37
live + death	16821	19856	16	−1.62

It is important to note that the log-likelihood score for each pair, rather than showing the actual strength of collocation, indicates the amount of evidence we have for rejecting the hypothesis of no association between the two words. A score of 3.84 or higher is significant at the 0.05 level, and a score of 6.63 is significant at the 0.01 level. In other words, a score of 6.63 indicates that the likelihood of this result being arrived at by chance is less than 1 in 100. The double line in the table splits the list into those pairs that are convincingly shown to be collocates (those above the line) and those that are not, assuming a threshold of 6.63.

The information collected in this study makes it possible to perform a similar calculation, this time not for the association between pairs of individual forms,

[7] The log-likelihood scores shown here were calculated using BNCweb (Hoffmann & Evert 2006), with a window of 4 words on either side of the node. Because the option of stopping at sentence boundaries was selected, the average number of words examined around each token is somewhat less than 8 (in fact, approximately 6.9).

but instead for the association between the entire family LIVE and the entire family DIE. This can be done by summing the total occurrences of all surface forms representing any of the 13 lemmas in LIVE, doing the same for the 6 lemmas in DIE, counting the number of co-occurrences of any of these in the corpus, and calculating the log-likelihood score for this co-occurrence. The result is shown in Table 3.

Tab. 3: Frequencies and log-likelihood scores for the families LIVE and DIE.

A + B	Freq. of A	Freq. of B	Co-occurrence of A + B	Log-likelihood score
LIVE + DIE	111912	56428	1406	1383.05

The log-likelihood value for LIVE + DIE is 1383, which is higher than 1263 (the value for *life* + *death*); this means that there is more evidence in the corpus for an association between these two families than there is for one between any single pair of words representing these families. This is logical and is a direct result of the method employed, which aggregates the data for all of the instances found of co-occurrence of the two families. The suggestion made here is that this latter method may possibly represent more faithfully what happens in the mental lexicon. Using both methods provides two different possibilities for modelling collocation as a psychological association.

Up to a point, the two methods yield the same results. For all of the pairs in Table 2 above the double line, there is sufficient evidence to point to a significant relation between them and thus to call the pairs collocates, so the two methods are in agreement. The difference begins when we reach the double line: according to the form-form calculation, there is no reason to claim an association between the pairs below this line, such as *alive* + *die*. By contrast, the family collocation result suggests an association between the two families that applies by default to all of the forms in both families. Importantly, this does not mean that all of the pairs of forms representing the two families are expected to have an equally strong relation – it merely makes the claim that there is *some* relation between any two forms from the two families. The combination of collocational tests at the word level and at the family level may therefore be seen as providing two different types of information, both of which are potentially useful to linguistic investigation.

If we repeat the comparison with the families WORK and EMPLOY, we arrive at a slightly different result. Table 4 shows the log-likelihood scores for all of the unique pairs of forms (ignoring part of speech) listed in (33) above. As can be seen, with a threshold of 6.63, eleven of the pairs are above this cutoff, and six of the pairs are below it. The highest log-likelihood value is 123.

Tab. 4: Frequencies and log-likelihood scores for pairs of surface forms representing the families WORK and EMPLOY.

A + B	Freq. of A	Freq. of B	Co-occurrence of A + B	Log-likelihood score
work + unemployed	89319	2761	77	122.89
worker + unemployed	3593	2761	15	66.04
worker + employer	3593	3002	14	57.55
worker + employed	3593	5023	16	54.28
working + unemployed	28636	2761	27	45.73
work + employment	89319	10620	109	31.12
work + employee	89319	3092	41	22.47
worker + employment	3593	10620	13	22.24
working + employee	28636	3092	20	21.20
work + employer	89319	3002	34	12.87
worker + employ	3593	1706	3	6.96
work + employ	89319	1706	18	5.59
working + employment	28636	10620	30	4.43
working + employ	28636	1706	3	-0.01
worker + unemployment	3593	6401	1	-0.18
works + employ	14132	1706	1	-0.21
work + unemployment	89319	6401	33	-0.37

However, the log-likelihood score for the families WORK and EMPLOY (shown in Table 5) is as high as 907, which, compared to 123 (the value for *work + unemployed*), shows that aggregating the data provides far greater support for an association than is found for any single pair. In theory, two families could exist for which the aggregate data, spread over a large number of members, could yield a significant result for those two families *even if none of the pairs themselves reached significance*.

Tab. 5: Frequencies and log-likelihood scores for the families WORK and EMPLOY.

A + B	Freq. of A	Freq. of B	Co-occurrence of A + B	Log-likelihood score
WORK + EMPLOY	164241	44253	1729	906.53

Finally, we may perform the same comparison with NATION and LEAD. Table 6 shows the log-likelihood scores for the thirteen most common pairs of forms representing these families. Again, with a threshold of 6.63, seven of the pairs make the cut-off (one of them just barely), and six do not.

Tab. 6: Frequencies and log-likelihood scores for pairs of surface forms representing the families NATION and LEAD.

A + B	Freq. of A	Freq. of B	Co-occurrence of A + B	Log-likelihood score
national + leader	37561	9159	132	249.13
nationalist + leader	1390	9159	28	142.12
national + leadership	37561	4734	53	78.28
international + leading	22113	11155	44	35.20
nation + leader	4322	9159	12	18.06
national + leading	37561	11155	42	6.65
nationally + lead	859	14325	4	6.63
international + leadership	22113	4734	22	0.30
international + lead	22113	14325	1	0.24
nationalist + lead	1390	14325	33	−0.08
national + lead	37561	14325	3	−0.15
nation + lead	4322	14325	10	−0.29
international + leader	22113	9159	132	−0.60

Again, the log-likelihood score for the families NATION and LEAD (shown in Table 7) is 473, higher than the highest score for any individual pair, at 249. This means that once again, there is evidence of an association between the families as wholes that contradicts the results for several of the pairs of forms. For example, while *national* + *leader* has a very high log-likelihood score, *international* + *leader* shows no evidence of being a collocation using the traditional method. Might it nevertheless be the case that the existence of the many other pairs of collocates in the list creates a psychological association between these words that is stronger than the textual evidence would indicate? We have no answer to this question as yet, but the possibility would seem to be worth investigating further.

Tab. 7: Frequencies and log-likelihood scores for the families NATION and LEAD.

A + B	Freq.of A	Freq of B	Co-occurrence of A + B	Log-likelihood score
NATION + LEAD	76598	65507	837	472.79

A final point worth noting concerns the tendency of families as wholes to have higher log-likelihood scores than their individual members; although we have seen this happen three times, it will not always be the case. If several of the lemmas in two families had no relation to each other, the overall score for the families could be lower rather than higher. For example, while the individual words *legal + pleasures* have a log-likelihood score of 0.41, the families LEGAL and PLEASE (with 15 lemmas each and a diversity score of 0.009) have the low score of −47.99. This indicates that these two families are not likely to co-occur.

7 Conclusion

This chapter has presented evidence of a phenomenon termed 'family collocation', which bears on the idea that collocation is a lexical relation in text that in some way mirrors lexical associations in the mind. It has been suggested that the common practice of examining the collocational behaviour of individual word forms in corpora may lead us to underestimate the extent of the networks of lexical associations that exist in the mind. Words that are paradigmatically, etymologically, or otherwise related to each other may share and reinforce lexical associations, which in turn can be expressed via a variety of surface forms. For example, if a combination of words such as *satisfactory conclusion* occurs frequently, then combinations such as *unsatisfactory conclusion* and *satisfying conclusion* are likely to benefit invisibly (from the point of view of the text) from this association. This study represents a first attempt to investigate the likelihood that this is, in fact, the case.

The chapter investigated the associations between 100 high-frequency families in the BNC, finding that the members of these families tend to co-occur on average in 13 different ways – that is, for any two families, an average of 13 unique combinations of lemmas were found. While in some cases, this number was as low as 1, in others it was high as 101. Overall, for any given pair of families, 21% of all possible combinations were found to occur.

A closer look at three case studies, involving the pairs LIVE + DIE, WORK + EMPLOY, and NATION + LEAD, indicated that often, the log-likelihood score for two families is higher than it is for any pair of members. This suggests that the co-occurrence of the various lemmas, in the aggregate, can provide evidence of an association that is missed when considering the individual forms. This association between families may potentially provide a psychological “boost” to the associations between less commonly co-occurring members.

Of course, we do not know what actually happens in the mind when such combinations are produced or perceived. It would be interesting to address this question through the use of psycholinguistic experimental methods. Another aspect of family collocation that has not been considered here is whether this relationship is likely to be symmetrical or asymmetrical. This is a general question about collocation which deserves much more study.

Another desirable step in attempting to determine what the significance of family collocation might be is to perform a large-scale, quantitative study in which an entire corpus is processed into families and the interactions among them are analysed. Although this falls well outside the scope of this exploratory study, it would increase the likelihood of the project begun here being concluded satisfactorily.

References

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh, UK: Edinburgh University Press.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 254–279.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London, UK: Pearson Education.

Biber, D., Conrad, S., & Cortes, V. (2004). ‘If you look at …’: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Burnard, L. (2007). Reference Guide for the British National Corpus (XML Edition). Online publication for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. Retrieved from http://www.natcorp.ox.ac.uk/docs/URG/index.html (last accessed October 2018).

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*(4), 411–33.

Cheng, W., Greaves, C., Sinclair, J. M., & Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*, *30*(2), 236–252.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly, 34*(2), 213–238.

Dąbrowska, E. (2014). Words that go together: Measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon, 9*(3), 401–418.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory, 5*(1), 61–78.

Garretson, G. (2010). *Corpus-derived Profiles: A Framework for Studying Word Meaning in Text.* Doctoral dissertation, Boston University, Boston, MA.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102–121). London, UK: Longman.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An Interdisciplinary Perspective*. Amsterdam, Netherlands: John Benjamins.

Gries, S. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 3–25). Amsterdam, Netherlands: John Benjamins.

Gries, S., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*(4), 635–676.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robbins (Eds.), *In Memory of J. R. Firth*. London, UK: Longmans.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London, UK: Routledge.

Hoffmann, S., & Evert, S. (2006). BNCweb (CQP-edition): The marriage of two corpus tools. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods* (pp. 177–195). Frankfurt am Main, Germany: Peter Lang.

Justeson, J. S., & Katz, S. M. (1992). Redefining antonymy: The textual structure of a semantic relation. *Literary and Linguistic Computing, 7*(3), 176–184.

Kövecses, Z. (1986). *Metaphors of Anger, Pride and Love: A Lexical Approach to the Structure of Concepts*. Amsterdam, Netherlands: John Benjamins.

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, *5*(2), 175–200.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam, Netherlands: John Benjamins.

Nordquist, D. (2004). Comparing elicited data and corpora. In M. Achard & S. Kemmer (Eds.), *Language, Culture and Mind* (pp. 211–223). Stanford, CA: CSLI.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh, UK: Edinburgh University Press.

Palmer, F. R. (1968). *Selected papers of JR Firth 1952-59*. Harlow, UK: Longmans.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam, Netherlands: John Benjamins.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. In Proceedings of the Beyond Named Entity Recognition Semantic Labelling for NLP tasks

Workshop tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004 (pp. 7—12). Lisbon, Portugal.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 127–151). Amsterdam, Netherlands: John Benjamins.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford, UK: Oxford University Press.

Sinclair, J. M. (2004). *Trust the Text: Language, Corpus and Discourse*. London, UK: Routledge.

Stubbs, M. (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, UK: Blackwell.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32, 231–254.

Wulff, S., & Titone, D. A. (Eds.). (2014). Bridging the methodological divide: Linguistic and psycholinguistic approaches to formulaic language [Special issue]. *The Mental Lexicon*, *9*(3).

Hildegunn Dirdal

Factors influencing the translation of *-ing* participial free adjuncts

Semantic role, context and a translator's individual style

Abstract: Previous studies have shown that independent, coordinated and subordinated clauses, as well as different kinds of phrases, are used to render *-ing* participial free adjuncts in translation from English to Norwegian. Some attention has been given to factors influencing the choice between these structures. The present paper reviews the factors investigated so far, and presents data from the Multiple-translation Corpus (an extension of the English–Norwegian Parallel Corpus) that make it possible to add several additional factors, related to the semantic role of the adjunct, the presence of coordination in the source sentence and the meaning and structure of the adjunct itself. The meaning and structure may be such that the Norwegian present participle can be used, in which case other solutions are dispreferred. The data also show that a translator's individual style is a contributing factor, influencing the choice between coordinate and subordinate clauses and the omission of coordinating conjunctions.

1 Introduction

English *-ing* participial free adjuncts pose a challenge for anyone translating into Norwegian, in which the use of the present participle is much more restricted. Previous studies have offered valuable insight into the translation of such adjuncts: which constructions are used, their relative frequencies and factors that influence the choice of construction. However, although several studies remark on the existence of individual variation, this is not something that has been studied in its own right. The present paper presents data on such variation, drawn from the Multiple-translation Corpus, an extension of the English–Norwegian Parallel Corpus which contains different translations of the same

Hildegunn Dirdal, University of Oslo, hildegunn.dirdal@ilos.uio.no

https://doi.org/10.1515/9783110489071-008

source texts.[1] The aim is to find out whether individual style influences the translation of *-ing* participial free adjuncts and to identify additional contextual factors.

The article starts with a background section defining the present participial free adjunct and giving an overview of previous research on the translation of such adjuncts from English into Norwegian (Section 2). Section 3 states the aims of the study in more detail and situates them in the context of previous findings, Section 4 describes the method and material, and Section 5 presents and discusses the results.

2 Background

Although both English and Norwegian have present participial free adjuncts, they are not used to the same extent. This background section starts with a definition of the present participial construction and a comparison of this construction in English and Norwegian (Section 2.1). It then presents other constructions that translators have recourse to when rendering *-ing* participial free adjuncts into Norwegian (Section 2.2), and, finally, gives an overview of the factors that previous studies have found to influence the choice of construction (Section 2.3).

2.1 English and Norwegian present participial free adjuncts

The construction in focus in this paper is the English present participial free adjunct. This construction is a non-obligatory constituent in a clause (i.e. an adjunct) and is headed by a non-finite verb in the present participle form, as in examples (1) and (2). The absence of an overt subject distinguishes it from the absolute construction exemplified in (3). Both adjuncts and absolutes can be augmented with subordinators (examples (4) and (5)), but these are restricted to *with*, *without* and *what with* for absolutes (Kortmann 1991: 5–8). I will limit my investigation to unaugmented (or bare) present participial free adjuncts (which are the most frequent), but will for the most part be using the shorter terms ‘present participial adjunct’, ‘*-ing* participial adjunct’ or simply ‘*-ing* adjunct’.

[1] For further information about the Multiple-translation Corpus, see the project website: http://www.hf.uio.no/ilos/english/services/omc/enpc/multtrans.html. The corpus is also described in Johansson (2004: 30–31).

(1) *Driving to Chicago that night*, I was struck by a sudden thought. (Quirk et al. 1985: 1121)

(2) Peter left *smiling*. (Haug et al. 2012: 162)

(3) *The coach being crowded*, Fred had to stand. (Kortmann 1991: 5)

(4) *Before returning home*, she bought presents for her parents. (Kortmann 1991: 8)

(5) *Without anyone noticing*, I slipped out of the room. (Quirk et al. 1985: 1121)

Present participial adjuncts are usually said to have adverbial functions, although those without dependents (such as *smiling* in example (2)) are sometimes classified as '(depictive) secondary predicates' (Fabricius-Hansen & Haug 2012: 3). Kortmann (1991: 119–121) proposes a scale for the most common semantic relations holding between present participial free adjuncts and their matrix clauses in his corpus, based on the amount of knowledge needed to arrive at the specific relation when interpreting the sentence. When less knowledge or contextual support is needed, Kortmann calls the relation 'less informative' or 'weaker', and when more knowledge or support is needed, he calls it 'more informative' or 'stronger'. He also draws a line between the less informative ones, which involve temporal overlap, and the more informative ones, which do not. Figure 9.1 in Kortmann (1991: 121) shows the semantic relations arranged along a scale of informativeness. Starting with the least informative relation, we find addition, accompanying circumstance, same time (simultaneity/overlap), exemplification/specification, and manner (all of which involve simultaneity), followed by time before (anteriority) and time after (posteriority), cause and result, instrument and purpose, condition, contrast and, finally, concession (neither of which involves simultaneity).

Free adjuncts can occur in initial, medial or final position in a matrix clause. In medial position, following the matrix subject, they can have either an adverbial or an adnominal function (non-restrictive modification of a noun) (Haug et al. 2012: 139; Kortmann 1991: 9). I will focus on adverbial uses of adjuncts in this paper, and exclude *-ing* adjuncts that are clearly part of noun phrases, as well as the one adjunct in my corpus that is ambiguous between an adnominal and an adverbial reading (see Section 4).

There are two written standards in Norwegian: Bokmål and Nynorsk. In both standards, a present participle can be formed by adding a suffix to the verb. The suffix is *-ende* in Bokmål and *-ande* in Nynorsk. Since the material for the present study consists of translations into Bokmål, this standard will be in focus below.

Norwegian present participial adjuncts have a more restricted internal syntax than English ones. According to Behrens, Fabricius-Hansen and Solfjeld (2012: 223), they cannot contain objects or reflexives. Although Kinn (2014: 71) argues that these elements cannot be completely ruled out, he acknowledges that they are infrequent. Fuhre (2010: 2) did not find a single instance in his material from the English–Norwegian Parallel Corpus. Elements that are more common in present participle clauses are adverbials and prepositional objects[2] (Kinn 2014: 73), illustrated in (6) and (7) respectively. Certain transitive verbs that can occur with prepositional objects are much less acceptable with regular objects in these present participial adjuncts. For the sake of illustration, such an example has been constructed on the basis of (7) and displayed in (8) below.

(6) ... jeg scootrer med vinden i håret [...] nedover mot Tøyen,
nynnende muntert med iPoden i øret ... (sites.google.com)
(Kinn 2014: 74)
Gloss: ... I scooter-PRES with wind-DEF in hair-DEF [...] down towards Tøyen, *hum-PRES.PART merrily with iPod-DEF in ear-DEF* ...[3]
'I am scootering with the wind in my hair ... down towards Tøyen,
humming merrily with my iPod in my ears ...'

(7) Kanskje kommer de tilbake til skolen *dansende på trinn fra* Giselle *eller nynnende på en strofe fra* Tryllefløyten?
(Oslo kommune, Utdanningsetaten 2015/16: 26)
Gloss: Maybe come-PRES they back from school *dance-PRES.PART on steps from* Giselle *or hum-PRES-PART on a tune from* The Magic Flute?

[2] Predicatives can also occur in present participle clauses, but are more common in those that function as different types of verbal complement, e.g. after the auxiliary BLI ('remain').

(i). De ble stående musestille.
They remain-PAST stand-PRES.PART mouse-still.
'They continued standing still as mice.'

This construction expresses a continuative or ingressive aspect (Faarlund, Lie & Vannebo 1997: 735; Kinn 2014: 76–77).

[3] The glosses follow the Norwegian examples word for word. Since Norwegian has definite endings, only sometimes in combination with a free-form article, the suffix gloss '-DEF' has been attached to definite nouns. In some cases verb endings are also glossed, but in the examples from the translation corpus used in the study, verbs are generally glossed with the equivalent forms in English. A paraphrase is added to Norwegian examples such as this one, but not in the examples from the translation corpus, where the original English sentence makes it superfluous.

‘Maybe they will come back from school *dancing steps from* Giselle *or humming a tune from* The Magic Flute?’

(8) * *dansende trinn fra* Giselle *eller nynnende en strofe fra* Tryllefløyten
Gloss: *danse-*PRES.PART *steps from* Giselle *or hum-*PRES-PART *a tune from* The Magic Flute
‘*dancing steps from* Giselle *or humming a tune from* The Magic Flute’

Most commonly, the present participle has no dependents, as in (9). It describes the subject rather than the action/event of the matrix clause, i.e. functions as a depictive or a free predicative rather than an event-oriented adverbial. In these cases, it is difficult to distinguish participles from adjectives, and they may also be coordinated with adjectives, as in (10).

(9) *Smilende* ga han hatten tilbake … (Olsen 2006)
Gloss: *Smile-*PRES.PART give-PAST he hat-DEF back …
‘*Smiling*, he handed the hat back …’

(10) Tenk å utsette en gammel dame for et slikt sjokk, sa hun *smilende og glad* på kirketrappa etter vielsen. (Isachsen 2012)
Gloss: Think to expose an old woman for a such shock, said she *smile-*PRES.PART *and happy* on church-stairs-DEF after (wedding-)ceremony-DEF
‘Imagine giving an old lady such a shock, she said, *smiling and happy*, on the church steps after the wedding ceremony.’

English present participial adjuncts can be both perfective and imperfective, whereas Norwegian present participial adjuncts are always imperfective and express simultaneity (Haug et al. 2012: 154, 176). They are stative or express unbounded activities (Behrens, Fabricius-Hansen, & Solfjeld 2012: 224). Not only does this mean that Norwegian present participle clauses have a narrower range of verbs, but these restrictions also limit the adverbial relations that present participial adjuncts can have to their matrix clause. All the non-co-eventive relations are ruled out, i.e. the more informative relations in Kortmann’s (1991: 121) scale, which do not involve temporal overlap: anteriority, posteriority, cause, result, instrument, purpose, condition, contrast and concession.

2.2 Translation alternatives

As the discussion in Section 2.1 has shown, Norwegian present participle clauses are much more restricted than English ones with respect to both form and meaning. As a consequence, Norwegian translators frequently have to resort to other structures in their rendering of English *-ing* adjuncts. Previous studies have shown that such adjuncts are most often rendered with coordination, followed by subordinate clauses and independent clauses, with different kinds of phrases and participle clauses forming smaller portions of the solutions (Fuhre 2010: 29; Johansson & Lysvåg 1987: 293; Smith 2004: 84). These four alternatives are illustrated in (11) with a sentence from the corpus used in this study, where the ten translators chose a range of different constructions. (The translators in the corpus are numbered and referred to as T1, T2, etc.)

(11) Source text: Bernard stalked behind the young men, *admonishing them.*

Alternative solutions:

Coordination:	T7: ... *og formante dem.* Gloss: ... *and admonished them.*
Subordinate clause:	T8: ... *mens han delte ut formaninger.* Gloss: ... *while he handed out admonitions.*
Independent clause:	Constructed:[4] ... *Han formante dem.* Gloss: ... *He admonished them.*
Phrase:	T6: ... *med kritiske påminnelser.* Gloss: ... *with critical reminders.*

[4] None of the ten translators used the solution of an independent clause as a translation of the adjunct in this particular case, but such solutions are found in other cases, for example in the following, where the independent clause is separated from the original matrix clause by a semicolon:

(ii). "We're going to be married," said Raymond, looking surprised, as though he himself had not known this until he said it.
"Vi skal gifte oss," sa Raymond; han så forbauset ut, som om han ikke hadde visst det selv før han sa det. (Translator 2)
Lit.: "We shall marry ourselves," said Raymond; he looked surprised out, as if he not had known it himself before he said it.

Both subordinate clauses and phrases may be of different kinds, and in the discussion of the data in Section 5, a more detailed categorization will be used. This will make it possible to investigate factors that influence more specific choices. In Fuhre's (2010: 34) material, finite adverbial clauses are the most frequent subordinate clauses found, followed by relative clauses, infinitive clauses and locative-relative clauses. When it comes to phrases, prepositional phrases are the most common, with smaller portions of adjective phrases, adverb phrases and noun phrases. Fuhre (2010: 35–36) chooses to treat present participles together with phrases. I treat them separately in order to explore factors affecting their use.

2.3 Factors influencing the choice of structure

There are several factors influencing the choice between the structures described above. Previous studies have explored some of these. The factor that has been investigated in most detail is the semantic relation between the *-ing* adjunct and the matrix clause. In addition, the situation types depicted by the clauses play a role, and there is some indication that genre may influence the choice. The following paragraphs will give a short account of the main findings with regard to these factors.

When a subordinate clause is chosen in the translation of an *-ing* participial free adjunct, the semantic relation it has to its matrix clause obviously plays a role in the choice of subordinator, which explicates this relation. However, even the choice between the four alternative constructions illustrated in (11) above is influenced by relation type. Fuhre (2010) surveys 682 *-ing* participial free adjuncts from the fiction part of the English–Norwegian Parallel Corpus, 442 of them from English originals translated into Norwegian. One of his aims is to investigate the correspondence types for the different semantic relations identified by Kortmann (1991: 121) (see Section 2.1). There are no examples of simultaneity, instrument or manner in his data, and relatively few examples of contrast, concession, condition and purpose. However, clear differences can be seen in the comparison of the numbers for accompanying circumstance, exemplification/specification, which Fuhre (e.g. 2010: 44, 50–54) calls 'elaboration', anteriority, cause and posteriority/result (treated together).

Tab. 1: Correspondence types for different relations found in Fuhre (2010: 51–65).

	Acc. circumstance	**Elaboration**	**Anteriority**	**Cause**	**Posteriority/ Result**
Indep. clause	24 (17.9%)	**21 (33.3%)**	12 (11.8%)	9 (10.7%)*	4 (12.1%)
Coordination	**43 (32.1%)***	**24 (38.1%)**	**66 (64.7%)**	16 (19.0%)*	**23 (69.7%)**
Subord. clause	**41 (30.6%)**	4 (6.3%)	16 (15.7%)	**45 (53.6%)**	3 (9.1%)
Phrase/participle	23 (17.2%)	12 (19.0%)	2 (2.0%)	12 (14.3%)	2 (6.1%)
Other	(2.2%)	2 (3.2%)	6 (5.9%)	2 (2.4%)	1 (3.0%)
Total	134 (100.0%)	63 (99.9%)	102 (100.1%)	84 (100.0%)	33 (100.0%)

Note: The numbers marked with an asterisk are calculated based on the raw numbers reported and deviate somewhat from the percentages given in Fuhre (2010).

Table 1 shows that translators strongly favour coordination and disfavour phrases when rendering *-ing* adjuncts expressing anteriority, posteriority and result. The relation of cause most often leads to the choice of a subordinate clause. Adjuncts expressing an accompanying circumstance give the most even spread of solutions, with coordination and subordinate clauses accounting for about a third each. With elaboration, it is independent clauses and coordination that are most strongly represented.

Studies of translation into Catalan and German confirm that meaning relation is an important factor, although the structural choices may be different. Even though Catalan does have a participle clause similar to the English *-ing* clause, Espunya (2009) finds that different renderings depend not only on the main types of meaning relation described above, but on subtypes within these. Doherty (1999) suggests that a distinction between 'foregrounding and backgrounding relations' can explain choices between phrasal and clausal translations in German.

In a study of the relation of accompanying circumstance, Behrens and Fabricius-Hansen (2005: 2–4) suggest that the combination of situation types expressed by the *-ing* adjunct and its matrix plays a role in the choice of German and Norwegian correspondences. They study correspondences from translation in both directions, i.e. both from and into English. Below, I consider the English–Norwegian correspondences in their examples (2) and (4)–(6), here quoted as (12)–(15), in order to show how different translations may be needed to express the same semantic relation when the situation types vary (the glosses are from Behrens and Fabricius-Hansen 2005: 2–4).

(12) Source text: The others followed her, *waving their weapons*.

Trans.: De andre kom etter *og viftet med våpnene sine*.
Gloss: The others came after *and waved with their weapons*.

(13) Source text: He saw Sandra off the premises, closed the shop, tidied up his office and left, *taking the four primary stones with him*.
Trans.: Han fulgte Sandra til døren, stengte butikken, ryddet opp på kontoret og gikk; *de fire kostelige steinene hadde han med seg*.
Gloss: … and left; *the four precious stones had he with him*.

(14) Source text: Da ble Sikita så glad at hun begynte å danse over engene *mens hun sang en sang hun hadde diktet i det kalde fengselet*.
Gloss of source text: … *while she sang a song* …
Trans.: Sikita was so happy that she began to dance across the meadow, *singing a song she had composed inside the dank prison*.

(15) Source text: He started to make a series of phone calls, *setting up getting-to-know-you meetings with the security chiefs in each of the main ministries*.
Trans.: Han satte i gang med en rekke telefonsamtaler *og avtalte møter "for å bli kjent med hverandre" med sjefene for sikkerhetsgruppene i de tre viktigste bygningene*.
Gloss: He started a series of phone calls *and set up getting-to-know-you meetings …*

In (12), the *-ing* adjunct expresses an activity which is co-temporal to the activity expressed in the matrix clause. The Norwegian translator has chosen coordination, which can also link two co-temporal activities. Example (13) is different in that the verbs in both the matrix and the *-ing* adjunct (*left* and *taking*) signal accomplishments. In this case, the translator has chosen an independent clause with a state verb to render the *-ing* adjunct (Behrens & Fabricius-Hansen 2005: 3).

Examples (14) and (15) have inchoative predicates in the matrixes. Behrens and Fabricius-Hansen (2005: 4) claim that although *sing a song* codes an accomplishment, the non-finite form it receives in the *-ing* clause imposes an unbounded activity reading so that it can be seen to overlap in time with the matrix situation. For that reading to be possible in Norwegian, a subordinate clause with the subordinator *mens* ('while') is needed. In (15), on the other hand, the addition of the plural *møter* ('meetings') to the accomplishment verb *avtale* ('set up') makes it unbounded, so that it is possible to achieve a co-temporal interpretation with simple coordination.

Behrens and Fabricius-Hansen (2005) do not explicitly say that coordination is impossible in (13), but Behrens, Fabricius-Hansen and Solfjeld (2012: 220), who compare the interpretation potential of non-finite adjuncts and their finite competitors, claim that coordination more naturally leads to a reading of temporal succession if the matrix clause is telic, whereas non-finite adjuncts, like that in the original sentence in (13), can "block a temporal succession interpretation".

Finally, there is some indication that genre may influence the choice of structure in Norwegian. Comparing the translation of initial *-ing* clauses in non-fiction and fiction, Smith (2004: 92–94) finds a larger proportion of adverbial clauses and prepositional phrases and a lower number of coordinated clauses in the former. However, as the two genres differ greatly with respect to the frequency of different semantic relations (e.g. a much lower number of *-ing* clauses expressing an accompanying circumstance in non-fiction), it is uncertain whether genre is a separate factor, and the matter will have to be investigated further.

3 The aims of the present study

Table 1 above shows that independent, coordinated and subordinate clauses, as well as phrases, are all possible translation alternatives for *-ing* adjuncts, although these constructions are used in different proportions for different semantic relations. This either means that there are further contextual factors that decide which construction is used, or that translators have a choice. It could also be a combination of the two.

Optionality would give scope for individual style to play a role. When translators have a choice, their individual preferences may lead to detectable patterns of usage (see e.g. Baker 2000: 245, 2004: 29; Marco 2004: 74; Munday 2008: 35). Dirdal (2014) shows that individual differences can be found in the way translators use clause building and clause reduction, which may thus be said to constitute part of a translator's voice. Her study focussed on the translations of English prepositional phrases and coordinate clauses into Norwegian, structures that have counterparts in the target language that are much more equivalent in form and function than what has been shown to be the case for present participial adjuncts (see Section 2.1). It would be reasonable to expect that individual style could also have an effect on the translation of *-ing* adjuncts.

The main aim of this study is thus to investigate to which extent a translator's individual style can be said to be an independent factor in the rendering of *-ing* clauses. The corpus used for this purpose contains translations by several translators of the same source text (see Section 4). Such a corpus makes it possible to keep individual style and context apart (they do not co-vary), and may therefore also allow identification of contextual factors that can be added to those described in Section 2.3, especially since the corpus is small enough for each instance to be examined in detail.

4 Material and method

The material for the present study comes from the Multiple-translation Corpus. This corpus was created in 1997 by Stig Johansson and Linn Øverås specifically for the study of individual variation between translators. It contains one fictional and one non-fictional English text, both of about 6000 words, and ten translations into Norwegian of each. Both texts were fairly recent at the time when the corpus was compiled and had not been translated before. The translators that were asked to contribute to the corpus were all experienced translators, and the ones translating the fictional text had also received prizes for their work (Johansson 2004: 30, 2011: 17).

Other translation corpora usually include translators who translated different source texts by different authors. This is the case for the corpora used in the studies referred to in Section 2.2 and 2.3. The Multiple-translation Corpus is unique in enabling comparison of translators while keeping constant the factors of source language, author's style, time period, etc.

For the present study, all the unaugmented -*ing* participial free adjuncts in the fictional text have been identified, and the ten translations have been compared in detail. The text is Byatt's (1996) short story 'A Lamia in the Cevennes'. The ten translators are anonymous, and are referred to by number (T1–T10).

The short story contains 31 clear cases of unaugmented -*ing* participial free adjuncts, excluding ambiguous cases (three in all): two instances can potentially be understood as noun phrases with *flashing* as an adjective premodifier, shown in example (16). Some translators seem to have interpreted them as -*ing* adjuncts, such as T1 in (16a), others as noun phrases, such as T2 in (16b). A further -*ing* clause (17) is ambiguous between an adnominal and an adverbial function. Since they are ambiguous, these three instances are not included in the analysis.

(16) Source text: In front of his prow or chin in the brightest lights moved a mesh of hexagonal threads, *flashing rainbow colours, flashing liquid silver-gilt, with a hint of molten glass* ...

a. T1: Når lyset var skarpest, blinket det foran baugen, eller altså haken hans, i et nett av heksagonale tråder, *det lynte i regnbuefarver, lynte i flytende sølvgyllent, med en antydning av smeltet glass ...*
Gloss: ... *it flashed in rainbow-colours, flashed in liquid silver-golden, with a hint of molten glass ...*

b. T2: Foran baugen eller haken hans i de sterkeste lysflekkene viftet et nett av heksagonale tråder *i glitrende regnbuefarger, i glitrende, flytende, forgylt sølv med en anelse smeltet glass ...*
Gloss: ... *in glittering rainbow-colours, in glittering, liquid, gilded silver with a hint of molten glass ...*

(17) The coil of pipe was uncoiled, the electricity was connected in his humming pumphouse, and a strange sound began, a regular boum-boum, like the beat of a giant heart, *echoing off the green mountain.*

When *-ing* clauses are part of existential constructions, as in (18), they are not considered to have the same adverbial role as the adjuncts dealt with in this paper (see Lee 2007: 165, where such clauses are listed as a separate class). These have therefore not been included in the analysis.

(18) And there was a head, *urging itself sinuously through the water beside his own...*

Out of the 31 unambiguous present participial free adjuncts, three are coordinated to other *-ing* adjuncts, as in (19). Only the first in each pair has been considered, since the translation of the second is not an independent matter, but depends on the choice for the first.

(19) She flung back her hair with an actressy gesture of her hands and sat down gracefully, *pulling the cheesecloth round her knees and staring down at her ankles.*

The total number of unaugmented *-ing* participial free adjuncts considered in the analysis is thus 28. One translator has omitted a section of text including one of these adjuncts. In all other cases, there are ten renderings of each, yielding 279 translations of *-ing* adjuncts altogether.

5 Results

Table 2 shows the constructions chosen by the ten translators. There are two types of variation – within and between translators. Every translator uses several strategies, but there are also different preferences between translators. It is this second kind of variation that may constitute individual style.

Tab. 2: Constructions chosen by the ten translators in their rendering of *-ing* participial free adjuncts[5].

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Total
Independent clause	3	2	1	1	1	1	1	1	1	1	13
Coordination	14	15	10	15	15	15	13	7	17	14	135
Finite adv. clause	3	5	9	5	6	4	8	12	4	6	62
That-clause				1	1	1		1			4
Relative clause								1			1
Pres. part. (clause)	2	4	4	2	3	6	4	3	4	2	34
Infinitive clause	1			1							2
Adjective phrase			1	1	1						3
Prepositional phrase	2	2	1	1	1	1	2	1	2	3	16
Adverb phrase			1							1	2
Noun phrase	1							1			2
Merged with matrix	1		1	1				1			4
Omission	1										1
Total	28	28	28	28	28	28	28	28	28	27	279

Table 2 shows that the preferred translation strategy for most of the translators is coordination, followed by finite adverbial clauses. However, there are also interesting differences. The most important one is that two of the translators (T3 and T8) use finite adverbial clauses more frequently than the others. This is clearly

[5] I have distinguished between the omission of an *-ing* clause when the rest of the sentence is translated and the non-translation of a whole portion of text, which might happen to include an *-ing* clause. T1 has an instance of the former and T10 an instance of the latter. T10 has thus been recorded as translating only 27 *-ing* clauses, and T1 as translating 28, but choosing the strategy of omission for one of them.

illustrated by (20), where T3 (20a) and T8 (20b) have chosen finite adverbial clauses with *idet* ('as') and *mens* ('while'), whereas all the other translators have chosen coordination, like T1 (20c).

(20) Source text: He swam even more than usual, *invoking the creature from time to time*.

a. T3: Han svømte til og med mer enn før, *idet han fra tid til annen påkalte dyret*.
Gloss: He swam to and with more than before, *as he from time to another invoked animal*-DEF.

b. T8: Nå svømte han enda mer enn før *mens han dann og vann påkalte skapningen*.
Gloss: Now swam he even more than before *while he now and again invoked creature*-DEF.

c. T1: Han svømte enda mer enn vanlig, *og påkalte vesenet fra tid til annen*. (Similar for T2, T4–7 and T9–10)
Gloss: He swam even more than usual, *and invoked creature*-DEF *from time to another*.

5.1 Semantic relation prevents coordination

Although coordination is the most frequent construction, it is not always a possible solution. When the *-ing* adjunct describes a time or condition for the event in the matrix clause, the most common translation is a finite adverbial clause, which all the translators used for the adjunct shown in (21), either one with a subordinating conjunction (*når*, 'when', in 21a) or one with inversion (21b). This solution is found in (22a) as well, but translators also used other constructions for the same adjunct, such as a merge of the *-ing* adjunct and the matrix clause (22b) or the use of an independent clause (22c) or a prepositional phrase (22d). The parentheses show which other translators used similar constructions.

(21) Source text: *Swimming in one direction*, he was headed towards a great rounded green mountain...

a. T1 (T2, T3, T4, T5, T6, T8, T9, T10): *Når han svømte den ene veien*, så han en stor rund grønn ås ...

Gloss: *When he swam the one way-DEF*, saw he a large round green hill...

b. T7: *Svømte han i den ene retningen*, var han på vei mot et stort, rundt og grønt fjell...
Gloss: *Swam he in the one direction-DEF*, was he on way towards a large, round and green mountain...

(22) Source text: Raymond made rather a noise *coming downstairs*.

a. T2 (T5, T7, T8): Raymond lagde et svare rabalder *da han kom ned*. [adverbial clause]
Gloss: Raymond made a tremendous racket *when he came down*.

b. T1: (T3, T4): *Raymond kom ikke direkte stille ned trappen.* [merge]
Raymond came not directly quietly down stairs-DEF.

c. T6: *Raymond kom ned*, det skjedde ikke nettopp lydløst. [independent clause + reordering]
Gloss: *Raymond came down*, it happened not exactly soundlessly.

d. T10 (T9): Raymond laget adskillig støy *på vei ned trappen*. [prepositional phrase]
Gloss: Raymond made considerable noise *on way down stairs-DEF*.

It does not seem possible to get a conditional reading like in (21) with coordination, except in imperatives like the hypothetical example (23), constructed on the basis of the sentence in (21). (This solution was obviously not adopted by any of the translators, as the imperative gives a different meaning.)

(23) Constructed ex.: Svøm i den ene retningen, og du vil se en stor, grønn ås...
Gloss: Swim in the one direction-DEF, and you will see a great, green hill...

One *-ing* adjunct, (24), is a specification of the event in the matrix clause rather than a reference to a separate action. In such cases, neither coordination with *og* ('and') nor a finite subordinate clause seems to be possible. Kortmann (1991: 110) also comments on the impossibility of adverbial clauses being alternatives to *-ing* adjuncts with a specifying role.

(24) Source text: And the colours changed as he watched them; the gold and silver lit up and went out, like lamps, the eyes expanded and contracted, the bars and stripes flamed with electric vermilion and crimson and then changed to purple, to blue, to green, *moving through the rainbow*.

a. T2 (T3, T5, T7): … og gled over i fiolett, blått og grønt, *i samme rekkefølge som regnbuen.*
Gloss: … and glided over in violet, blue and green, *in same order as rainbow-DEF.*

b. T1: … og så gikk de over i fiolett og blått og grønt**,** *alle regnbuens farver*.
Gloss: … and then went they over in violet and blue and green, *all rainbow-DEF-GEN colours*.

c. T8: … for så å gå over i purpur, blått, grønt, *hele veien gjennom regnbuen.*
Gloss: …for then to go over in purple, blue, green, *all way-DEF through rainbow-DEF.*

d. T10: … for så å gå over i purpur, i blått, i grønt, *og videre gjennom hele regnbuen.*
Gloss: … for then to go over in purple, in blue, in green, *and further through whole rainbow-DEF.*

e. T4 (T9): … for å gå over til purpur, til blått, til grønt, *beveget seg gjennom hele regnbuens spekter*.
Gloss: … for to go over to purple, to blue, to green, *moved themselves through whole rainbow-DEF-GEN spectrum.*

f. T6: … for så å gå over i purpur, i blått, i grønt, *pulserende gjennom hele regnbuen.*
Gloss: … for so to go over in purple, in blue, in green, *pulsating through whole rainbow-DEF.*

Example (24) shows that most of the translators chose a phrase in this case: a prepositional phrase (24a), noun phrase (24b and c) or adverb phrase (24d). However, some translators chose a clause without a subject, linked to the matrix clause only with a comma (24e). Although this type of construction has been labelled coordination, the lack of a coordinating conjunction is important. The clause/predicate presents an alternative description of the same event rather than one added to it. Fuhre (2010: 31) calls these “run-on sentences” and classifies them as independent clauses even when they share their subject with the matrix clause. He finds many of them with the semantic relation of specification, which he calls elaboration (e.g. Fuhre 2010: 50–51). However, as can be seen from Table 1, he also finds normal coordination to be frequent with this type of relation. The reason might be found in the way he categorizes his data (2010: 43): adjuncts are classified according to the strongest interpretation they can have on Kortmann’s scale (1991: 121), described in Section 2.1, where semantic roles are ordered according to how much knowledge or contextual support is needed to arrive at each interpretation. Since *-ing* clauses are often open to several interpretations, it might be that some of Fuhre’s elaborations have been interpreted as accompanying circumstances by translators, a relation lower on Kortmann’s scale. To be certain whether normal coordination is impossible with strict specification/elaboration, it would be necessary to look more closely at a larger amount of data.

One translator (T6) used a present participle clause in (24f) above. As described in Section 2.1, Norwegian present participial adjuncts are most commonly without dependents, but the dependents that may occur are adverbials and prepositional objects. T6 used a participle with an adverbial dependent. The other restrictions on present participial adjuncts also hold: it is co-eventive with the matrix clause and may be interpreted as expressing an unbounded activity.

The *-ing* adjunct in (25) can also be seen as specification, and was indeed translated in a similar way as (24) above by three of the translators (25a), in the sense that the *that*-clause they chose echoes and reformulates the last part of the previous stretch of the sentence, which also contains a *that*-clause. However, the same *-ing* adjunct can also be seen as a reason adverbial, and was translated with a finite adverbial clause by six translators (25b). One translator used a *that*-clause (25c), but also changed the order of the two clauses and inserted the adverb *således* (‘thus’) to mark the relationship between them.

(25) Source text: Bernard did not point out that he had not made a pact, *not having answered her request yes or no*.

a. T4 (T5, T8): Bernard lot være å peke på at han ikke hadde inngått noen pakt, *at han verken hadde svart ja eller nei på hennes anmodning*.
Gloss: Bernard let be to point at that he not had made any pact, *that he neither had answered yes or no on her request*.

b. T10 (T1, T2, T3, T7, T9): Bernard unnlot å påpeke at han ikke hadde inngått noen pakt, *siden han hverken hadde svart ja eller nei på spørsmålet hennes*.
Gloss: Bernard neglected to point-out that he not had made any pact, *since he neither had answered yes or no on question-*DEF *her*.

c. T6: Bernard påpekte ikke *at han hverken hadde sagt ja eller nei til hennes bønn*, og at det *således* ikke eksisterte noen pakt mellom dem.
Gloss: Bernard pointed-out not *that he neither had said yes or no to her plea*, and that there *thus* not existed any pact between them.

5.2 Contexts favouring Norwegian present participle renderings

There is a further group of *-ing* adjuncts, shown in (26)–(29), that were not translated with coordination. (There was one exception, but in that case the rendering of the *-ing* clause had been placed in front of the matrix, i.e. their order had been inverted.) The four adjuncts in this group were overwhelmingly translated with present participle clauses (31 out of 39 cases[6]). The second most frequent translation is a prepositional phrase. The avoidance of coordination can be explained by the temporal succession reading imposed by coordination when the first clause is telic (Behrens, Fabricius-Hansen, & Solfjeld 2012: 220, see Section 2.3). The eventualities in these four matrix clauses are achievements, and thus telic: *come out of the car*, *appear*, *run* (in the meaning of 'flee') and *begin to sink*. However, to explain the preference for present participles instead of other constructions, such as prepositional phrases or finite adverbial clauses, I will argue that the form and content of the sentences are such that some of the more common present participle constructions in Norwegian are possible. With respect to meaning, all the four *-ing* adjuncts are imperfective and co-eventive with the matrix clauses, both of which constitute requirements on Norwegian present participial adjuncts (see Section 2.1). With respect to form, they also fit well with the restrictions described in Section 2.1. First of all, the sentence in (26) has a one-word adjunct – *smiling*. The most common present participial free adjunct in Norwegian is one without dependents, describing the subject rather than the matrix

[6] This is where a section is missing from T10.

eventuality. Example (26) shows a sentence where all the translators have chosen this type of adjunct.

(26) Source text: Raymond Potter came out of the car *smiling* ...

T2 (+all the rest): Raymond Potter kom *smilende* ut av bilen ...
Gloss: Raymond Potter came *smiling* out of car-DEF ...
[Three translators have used the verb *steg* ('stepped') instead of *kom* ('came').]

In example (27), the adjunct is also short, and may be seen as participant-oriented rather than event-oriented. One of the translators chose a present participle without dependents (27a), which is the most natural-sounding in Norwegian. Prepositional phrases are also possible, such as the one used by T10 (27b), but again the information in the adverbial *still* is lost. A majority of translators seem to have wanted to keep this information. Adverbials being possible in Norwegian present participle clauses, this solution was favoured (five translators in 27c) over a finite adverbial clause (only one translator in 27d), perhaps due to the length of the latter and the medial position of the adjunct.

(27) Source text: The swimming-pool soughed and sighed and began *still sighing* to sink ...

a. T2: Bassenget sukket og stønnet og begynte *stønnende* å synke ...
Gloss: Pool-DEF sighed and groaned and began *groaning* to sink ...

b. T10: Svømmebassenget surklet og stønnet, og vannstanden begynte å synke *med en sukkende lyd* ...
Gloss: Swimmingpool-DEF gurgled and groaned, and water-level-DEF began to sink *with a sighing sound* ...

c. T3 (T5, T6, T7, T8): Svømmebassenget sukket og stønnet og ga seg, *fremdeles sukkende*, til å tømmes ...
Gloss: Swimmingpool-DEF sighed and groaned and gave itself, *still groaning*, to to be-emptied ...

d. T4: Bassenget suste og sukket og begynte *ennå mens det sukket* å synke ...
Gloss: Pool-DEF soughed and sighed and began *still while it sighed* to sink ...

An adverbial dependent is also found in the most frequent translation of the *-ing* adjunct in the third sentence in this group (28), whereas the most frequent translation of the adjunct in the final sentence has a prepositional object (29), which is also possible in Norwegian present participial adjuncts.

(28) Source text: Many men might have run *roaring in terror* ...

T2 (T3, T4, T6, T7, T8, T9): Mange menn kunne ha flyktet *brølende av skrekk* ...
Gloss: Many men could have fled *roaring of terror* ...

(29) Source text: The two young men appeared *carrying an immense boa-constrictor of heavy black plastic pipe* ...

T1 (T3, T5, T6, T7, T8, T10): De to unge mennene kom *bærende på en enorm boa constrictor i tykk sort plast* ...
Gloss: The two young men came *carrying on an enormous boa constrictor in thick black plastic* ...
[T8 has *kom slepende på* ('came dragging on'), T10 *kom slepende med* ('came dragging with') and T3 and T6 *dukket opp bærende på* ('emerged carrying on').]

In addition to the match with the internal syntax of Norwegian present participial adjuncts, it is possible that the similarity to one of the few common complement constructions with present participles in Norwegian plays a role in (29). Kinn (2014: 80–83) argues that KOMME ('come') may act as an auxiliary taking present participles of certain motion verbs as complements, as in (30), and that there is an overlap between this construction and constructions with KOMME as a main verb followed by a present participle subject predicative, as in (31).

(30) En mann som heter Thomas kom *kjørende i en diger lastebil.* (www.ringblad.no) (Kinn 2014: 80)
Gloss: A man who is-called Thomas come-PAST *drive-PRES.PART in a huge truck.*
'A man called Thomas came *driving (in/up) in a huge truck.*'

(31) Noen kom *sjanglende* inn fra kneipene langs "Ølklemmå" og "Subakanalen" ... (erlingjensen.net) (Kinn 2014: 82)
Gloss: Someone come-PAST *stagger-PRES.PART* in from joints-DEF along "Ølklemmå" and "Subakanalen" ...

'Someone came *staggering* in from the joints along "Ølklemmå" and "Subakanalen" ...'

The possibility of using *kom* ('came') to render *appeared* in (29), might thus create a favourable context for the use of a present participle in Norwegian.

5.3 Coordination dispreferred when the context already contains coordination

Even though coordination is the most common solution overall, there are some cases where all or almost all of the translators used subordinate clauses. These cases turn out to involve sentences where coordination is already used somewhere else. Example (32) shows a sentence where there is coordination of two predicates before the *-ing* adjunct. Nine of the ten translators chose to render this *-ing* adjunct as a finite adverbial clause, illustrated with the solution from T2 (32a). The only translator to render the *-ing* adjunct as a coordinate clause is T1, who omitted the coordinated predicate from the original and thus also avoided two coordinated clauses in a row (32b).

(32) Source text: Bernard turned over on his side, and floated, *disentangling his brown legs from the twining coloured coils*.

a. T2: Bernard snudde seg over på siden og fløt *mens han viklet de brune bena ut fra de omslyngende fargede kveilene*.
Gloss: Bernard turned himself over on side-DEF and floated *while he disentangled the brown legs-DEF from the twining coloured coils-DEF*.

b. T1: Bernard la seg over på siden *og frigjorde sine brune ben fra de farveglade buktningene*.
Gloss: Bernard laid himself over on side-DEF *and made his brown legs free from the colourful coils-DEF*.

The coordination in the original may also occur in the *-ing* adjunct itself, as illustrated in (33). Again, nine translators used a finite adverbial clause, similar to the one used by T9 (33a). The only exception is T10, who reordered the content and made the *-ing* adjunct into an independent clause (33b).

(33) Source text: Don't go, he begged it, *watching and learning*, don't go.

a. T9: Ikke forsvinn, tryglet han den *mens han så og lærte*, ikke forsvinn.
Gloss: Not disappear, he begged it *while he saw and learnt*, not disappear.

b. T10: *Han iakttok og lærte*. Ikke fly bort, ba han, ikke fly bort.
Gloss: *He watched and learnt*. Not fly away, he begged, not fly away.

5.4 Individual style

The clearest indication of individual differences found in Table 2 was in the relative frequency difference between coordination and subordination. To control for other factors, it was deemed important to compare these frequencies only with regard to sentences that could potentially be translated using both coordination and subordination.

Tab. 3: Translation of *-ing* participial free adjuncts when both coordination and subordination is possible.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Total
Independent clause	3	2	1	1	1		1	1	1		11
Coordination	14	15	10	14	14	15	13	7	16	14	132
Subordinate clause	1	1	5	2	2	2	3	9	1	3	29
Phrase			2	1	1	1	1	1		1	8
Total	18	18	18	18	18	18	18	18	18	18	180

The following cases were therefore excluded from the comparison in Table 3: cases where coordination could not be used because of the semantic relation expressed or where coordination was dispreferred because of a telic matrix clause, and cases where all the translators chose subordinate clauses because there was another case of coordination in the same sentence (see Sections 5.1 and 5.3). This leaves only sentences where there is a genuine choice between coordination and subordination (as well as other possibilities). The trends are still clear, with T3 and T8 favouring subordination much more frequently than the rest.

A further sign of personal style can be seen in the translators' decisions to leave out the conjunction *og* ('and') when coordination is used, i.e. when a predicate is added that shares its subject with the matrix clause. Table 4 shows how often the translators used coordinated clauses with or without a coordinating conjunction.

Tab. 4: Use of coordination with and without conjunction in the rendering of *-ing* participial free adjuncts.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Total
With conjunction	12	14	10	11	15	15	13	7	13	13	123
Without conjunction	2	1		4					4	1	12
Total	14	15	10	15	15	15	13	7	17	14	135

Sentence (24) above showed an example of conjunctionless coordination, or the attachment of a new predicate after a comma, where normal coordination was impossible because of the semantic relation of specification. Only T4 and T9 used this strategy. The same two translators chose a similar structure in the rendering of three other sentences, one of which is shown in (34). Only in one sentence, shown in (35), did as many as five translators choose the same solution. However, in this specific case, the clause in question can be seen as the penultimate in a coordinated list, where it is common to leave out the conjunction.

(34) Source text: He swam more and more, *trying to understand the blue …*

T4 (T9): Han svømte mer og mer, *forsøkte å forstå denne blåfargen …*
Gloss: He swam more and more, *tried to understand this blue-colour …*

(35) Source text: She flung back her hair with an actressy gesture of her hands and sat down gracefully, *pulling the cheesecloth round her knees* and staring down at her ankles.

T1 (T2, T4, T9, T10): Hun slengte håret tilbake med en filmstjerneaktig håndbevegelse og satte seg grasiøst, *trakk skinkeposestoffet over knærne* og tittet ned på anklene sine.
Gloss: She slung hair-DEF back with a film-star-like hand-movement and sat herself gracefully, *pulled cheesecloth-DEF over knees-DEF* and looked down at ankles-DEF her.

6 Conclusion

The present study adds new knowledge about several factors that influence the translation of English unaugmented *-ing* participial free adjuncts into Norwegian, in addition to confirming the importance of semantic role and situation type.

The data have revealed the same preference for coordination in the translation of *-ing* clauses as has been found in previous studies. The second most frequent structure is subordination, with independent clauses and phrases making up smaller portions of the renderings. The data confirm previous findings that the semantic relation between the *-ing* clause and its matrix plays a role in the choice of a Norwegian construction. This was seen mainly in two cases: Firstly, it was found that coordination is not possible when the *-ing* clause describes a time or condition for the event in the matrix clause. In these cases, finite adverbial clauses are preferred, the subordinator making the relation explicit. However, other solutions are also found, such as the use of prepositional phrases or independent clauses. Secondly, neither normal coordination with a coordinating conjunction nor finite adverbial clauses are used when the *-ing* adjunct expresses specification/explicitation. Besides prepositional phrases and other types of phrases, translators use coordination without a coordinating conjunction, or what has been called "run-on sentences" (Fuhre 2010: 53).

The results also confirm the role of the situation type. Present participial free adjuncts can be co-eventive with a telic matrix clause, but in these cases coordination will lead to a temporal sequence interpretation and is thus avoided.

In addition, it is found that coordination is dispreferred when there is other coordination in the near vicinity of the *-ing* adjunct, and, further, that some *-ing* adjuncts have a meaning, structure and context that favour the use of the Norwegian present participle. Co-eventive and imperfective adjuncts that consist only of a present participle and are participant-oriented, are very likely to be rendered as present participial adjuncts in Norwegian, but this also happens frequently if the dependents of the *-ing* participle can be expressed as adverbs and prepositional objects in Norwegian.

Finally, the data show clear individual differences between translators, indicating that a translator's style is also an influencing factor in the translation of *-ing* participial free adjuncts. Such individual differences were found in the choice between coordination and subordination (when both were possible) and in the use of coordinate predicates without a coordinating conjunction.

References

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target*, *12*(2), 241–266.

Baker, M. (2004). The treatment of variation in corpus-based translation studies. In K. Aijmer & H. Hasselgård (Eds.), *Translation and Corpora: Selected Papers from the Göteborg-Oslo Symposium 18-19 October 2003* (pp. 7–17). Göteborg, Sweden: Acta Universitatis Gothoburgensis.

Behrens, B. & Fabricius-Hansen, C. (2005). The relation Accompanying Circumstance across languages: Conflict between linguistic expression and discourse subordination? *SPRIKreports, 32*. Oslo, Norway: University of Oslo. Retrieved from http://www.hf.uio.no/ilos/forskning/prosjekter/sprik/pdf/bb/Sprik-Report32-bb-cfh.pdf (last accessed May 2017).

Behrens, B., Fabricius-Hansen, C., & Solfjeld, K. (2012). Competing structures: The discourse perspective. In C. Fabricius-Hansen & D. Haug (Eds.), *Big Events, Small Clauses: The Grammar of Elaboration* (pp. 179–225). Berlin, Germany: De Gruyter.

Byatt, A. S. (1996). A lamia in the Cevennes. In C. Hope & P. Porter (Eds.), *New Writing*, (Vol. 5, pp. 1–17). London, UK: Vintage.

Dirdal, H. (2014). Individual variation between translators in the use of clause building and clause reduction. *Oslo Studies in Language*, *6*(1), 119–142.

Doherty, M. (1999). The grammatical perspective of *-ing* adverbials and their translation into German. In H. Hasselgård & S. Oksefjell (Eds.) *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 269–282). Amsterdam, Netherlands: Rodopi.

Espunya, A. (2009, 9–12 September). *V-ing free adjuncts and implicit contingency relations in three science popularisation works and their translations*. Paper presented at the 42nd Annual Meeting of the Societas Linguistica Europaea, Workshop on Connectives Across Languages: Explicitation and Grammaticalization of Contingency Relations, Ghent, Belgium. Abstract retrieved from http://www.francais.ugent.be/workshop/preview (last accessed May 2017).

Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk Referansegrammatikk*. Oslo, Norway: Universitetsforlaget.

Fabricius-Hansen, C., & Haug, D. T. T. (2012). Introduction. In C. Fabricius-Hansen & D. Haug (Eds.), *Big Events, Small Clauses: The Grammar of Elaboration* (pp. 1–17). Berlin, Germany: De Gruyter.

Fuhre, P. (2010). *The English -ing Participial Free Adjunct in Original and Translated Fiction: An English–Norwegian Parallel Corpus Study*. Master's dissertation. University of Oslo, Oslo, Norway.

Haug, D. T. T., Fabricius-Hansen, C., Behrens, B., & Helland, H. P. (2012). Open adjuncts: Degrees of event integration. In C. Fabricius-Hansen & D. Haug (Eds.), *Big Events, Small Clauses: The Grammar of Elaboration* (pp. 131–178). Berlin, Germany: De Gruyter.

Isachsen, A. (2012). Australsk bryllup i Trondenes kirke. *Harstad Tidende* [Online newspaper]. Retrieved from http://www.ht.no/incoming/article574451.ece (last accessed May 2017).

Johansson, S. (2004). Why change the subject? On changes in subject selection in translation from English into Norwegian. *Target*, *16*(1), 29–52.

Johansson, S. (2011). Between Scylla and Charybdis: On individual variation in translation. *Languages in Contrast 11*(1), 3–19.

Johansson, S. & Lysvåg, P. (1987). *Understanding English Grammar, Part II: A Closer View*. Oslo, Norway: Universitetsforlaget.

Kinn, T. (2014). Verbalt presens partisipp. *Norsk Lingvistisk Tidsskrift*, *32*(1), 62–99.

Kortmann, B. (1991). *Free Adjuncts and Absolutes in English: Problems of Control and Interpretation*. London, UK: Routledge.

Lee, S.-A. (2007). *Ing* forms and the progressive puzzle: A construction-based approach to English progressives. *Journal of Linguistics*, *43*(1), 153–195.

Marco, J. (2004). Translating style and styles of translating: Henry James and Edgar Allan Poe in Catalan. *Language and Literature*, *13*(1), 73–90.

The Multiple-translation Corpus (1997–1999), Dept. of British and American Studies, University of Oslo (http://www.hf.uio.no/ilos/tjenester/kunnskap/sprak/omc/enpc/multtrans.html). Compiled by Linn Øverås and Stig Johansson in connection with the English–Norwegian Parallel Corpus. http://www.hf.uio.no/ilos/english/services/omc/enpc/ (Last accessed Dec. 2017).

Munday, J. (2008). *Style and Ideology in Translation: Latin American Writing in English*. New York, NY: Routledge.

Olsen, T. (2006). Galante Haakon hentet hatten. *Nettavisen* [Online newspaper]. Retrieved from http://www.nettavisen.no/nyheter/galante-haakon-hentet-hatten/723572.html (last accessed May 2017).

Oslo kommune, Utdanningsetaten (2015/16) Den kulturelle skolesekken 2015/16 - grunnskolen 1.–7. trinn. Retrieved from: https://www.dks.osloskolen.no/res/ck/files/oslo/2015-16/DKS_2015_1_7trinn_WEB.pdf (last accessed May 2017).

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London, UK: Longman.

Smith, M.-B. M. (2004). *Initial -ing Clauses in English and their Translation into Norwegian*. Master's disstertation. University of Oslo, Oslo, Norway.

Leonie Wiemeyer

The diachronic productivity of native combining forms in American English

Abstract: Combining forms (CFs) are bound lexical elements which are abundant in the English language, such as the well-established CFs *eco-*, *geo-*, *-holic*, and *-athon*. In the 20th century, a new class of so-called 'native combining forms' emerged. These often occur in jocular formations, predominantly on the Internet. They are used to name new cultural phenomena and have given rise to a considerable number of neologisms in the past two decades. This chapter presents the results of a corpus study of the productivity of native CFs in written American English. The productivity of 21 CFs was examined diachronically from 1950 until 2009 using the Corpus of Historical American English (COHA). The aim of the study was to determine whether the elements investigated are currently productive and how this productivity has changed over time. A further aim was to identify factors influencing productivity and to establish a connection between an element's productivity and its topicality.

1 Introduction

Combining forms are bound lexical elements which combine with words, other combining forms or affixes to create lexical words (Fradin 2000). Based on whether they occur word-initially or word-finally, they are divided into initial and final combining forms (Bauer 1983). Thus far, they have not been given much attention in the linguistic literature, but the large number of elements classified as combining forms in dictionaries such as the Oxford English Dictionary (OED) testifies to the fact that combining forms are not just a fad. As a morphological category, they are quite difficult to grasp. Combining forms are heterogeneous in form, origin, and syntactic function (see Bauer 1983; Warren 1990; Fradin 2000). They share properties with affixes and blends, among others, and the formations they give rise to resemble compounds (Fischer 1998). Linguists disagree on the categorisation of the elements in question as well as on their characteristics (see Warren 1990 for a survey of definitions).

Leonie Wiemeyer, University of Bremen, wiemeyer@uni-bremen.de

https://doi.org/10.1515/9783110489071-009

Before the 20th century, only Latinate and Ancient Greek elements, so-called neo-classical combining forms, played a role in the English vocabulary (see Bauer 1998). More recently, the relatively new class of native combining forms has gained importance. Native combining forms are derived from English lexical words. They are usually the result of a reinterpretation of the morphological make-up of a source word and the subsequent use of one of its parts as a bound element (Warren 1990). An example is the element *-thon*, which is not a morpheme in the source word *marathon*, but has been reinterpreted as such in English and gives rise to formations such as *walkathon* and *readathon* denoting long-lasting events, usually held for charity. These native elements are not borrowed from other languages, unlike neo-classical elements, which are also combining forms in Latin or in (Ancient) Greek (Lüdeling 2006). Their form and use is often very different from that of neo-classical combining forms. Like the neo-classical elements, they form so-called combinations (Fischer 1998), which are compound-like formations containing combining forms, e.g. *shopaholic* and *workaholic* from *-(a)holic*.

Native combining forms often occur in jocular formations or as attention-getters, predominantly on the Internet, in the media and in advertising. Because such coinages appear novel, unusual, and modern, they are likely to catch the reader's eye and be remembered, especially if their meaning is transparent (see Lehrer 2003 for similar observations regarding blends). This is exemplified by the underlined formations below:

(1) SkinnyLicious® is our collection of fresh and delicious menu options with lower calories and signature rich taste.[1]

(2) Give him space; he'll reward you later with a sexathon. (Corpus of Contemporary American English (COCA),[2] MAG: Cosmopolitan, 2011)

(3) A self-confessed "fishaholic," he makes time to get out at least weekly for trout or bass and manages yearly trips for bonefish and Atlantic salmon. (COCA, MAG: field and stream, 1992)

(4) Its Cookie-ception: Oreo Cookies Now Come In Chocolate Chip Flavours![3]

Native combining forms are a special type of morpheme because they add lexical meaning to free lexemes which cannot be adequately expressed by established

[1] Taken from https://www.thecheesecakefactory.com/menu/skinnylicious/, last accessed February 2017.

[2] A list of corpora and dictionaries cited appears at the end of this chapter.

[3] Taken from https://www.hungryforever.com/cookie-ception-oreo-cookies-now-come-chocolate-chip-flavours/, last accessed February 2017.

bound morphemes (see Bauer 1983). These morphemes have contributed a considerable number of unusual neologisms with novel meanings to the English language in the last few decades, many of which have become established, and new words are readily coined (see e.g. Fischer 1998). Speakers in a variety of contexts and media exploit their creative and innovative potential, as indicated by the examples above. Nevertheless, despite their remarkable contribution to the lexicon of the English language, very little is known about the productivity of such native elements and which factors play a role when a combining form becomes productive.

This chapter explores the characteristics of native combining forms, above all their productivity, in written American English. The following section reviews the issue of defining combining forms and delineating them from other morphological elements. It also provides a list of distinctive features of combining forms. Previous research on the productivity of native combining forms is surveyed in Section 3. In Section 4, the type frequencies of a set of native combining forms are surveyed diachronically from 1950 to 2009 on the basis of data from the Corpus of Historical American English (COHA; Davies 2010–) in order to establish whether the examined native combining forms are productive word-forming elements in written American English. Results are presented in Section 5. Finally, factors influencing the productivity of combining forms are considered in Section 6.

2 Defining the category of combining forms

Developing a definition of the category of combining forms as a basis for a study of their productivity is far from an easy task. The heterogeneity of the elements listed in the literature (see, for example, Fischer 1998; Warren 1990) and in dictionaries indicates that the notion 'combining form' is by no means clear-cut. There is generally very little agreement on how combining forms are to be defined, which elements can be classified as combining forms, and where to draw the line between combining forms and other word-forming elements such as affixes, stems, bound roots, and free lexemes. As Prćić asserts, "the labelling of all bound lexical elements in dictionaries, both pedagogical and native-speaker ones, is inconsistent and confusing, sometimes even contradictory and mutually exclusive" (2005: 314).

Combining forms are neither free lexemes nor affixes, but rather exist somewhere along a continuum between the two. The categorisation of possible members of the category 'combining form' – for example elements such as *hydro-*,

-crat, *-ology*, *-aholic*, *-scape* and *-friendly* – is often ambiguous.[4] Numerous terms can be found in the literature describing the same word-forming elements. They are referred to as pseudo-prefixes and -suffixes, semi-prefixes and -suffixes, quasi-affixes, prefixoids and suffixoids, blends or splinters (cf. Adams 1973: 188; Fischer 1998: 55; Schmid 2005: 47). However, these terms are often applied without a definition and thus appear to be more useful for the description of the elements' similarity to affixes or of the way they were generated than as categories of morphological elements. Fischer (1998) found that the treatment of combining forms varies between as well as within the references she consulted. Furthermore, she noticed that various rather vague descriptions of these elements were used in the OED where they were sometimes described as 'in combination', 'terminal element' or 'suffix or final element'. She concluded that "[t]he lack of certainty in the way such forms are evaluated by the lexicographers is clearly shown here. It is not clarified whether these forms are affixes, compositional elements, or something in between" (Fischer 1998: 58).

This uncertainty in the classification is perpetuated in the treatment of most of these bound lexical elements. The word-forming element *-aholic/-(o)holic*[5] is a palpable example. While it is listed as a noun-forming suffix in the OED, the *Macmillan Dictionary* and the *Longman Dictionary of Contemporary English*, it is defined as a '(noun) combining form' in the online dictionary of Merriam-Webster. Warren defines *-aholic* as a combining form created by secretion, which is "the result of some folk-etymological misdivision of morphemes (i.e., *alcoholic* would mistakenly be believed to consist of *alco* + *holic*)" (Warren 1990: 117). Lehrer (1998: 3) analyses the element as a combining form that has resulted from blending of the base words, for example *chocolate* + *alcoholic*, while Kolin (1979) and Algeo (1981) consider it a pseudo-suffix. Schmid (2005: 170) describes it as a 'suffix-like element' created by secretion, and Fradin treats *-aholic* as a prototypical example of what he calls 'secreted affixes' (cf. 2000: 46).

According to Kastovsky (2009: 3), the problem of properly defining and categorising combining forms lies in the fact that "the linguistic status of these elements was never really made clear […], nor were there any criteria by means of which they could be distinguished from other types of lexical element such as words, roots, stems or affixes". This problem still exists today. In many cases,

[4] See Lüdeling (2006) for an overview of the conflicting viewpoints on the status of neo-classical elements.

[5] All three forms are attested; *-oholic* and *-holic* are listed as variants of *-aholic* in the OED. The spelling of the linking vowel appears to be a matter of taste; the variant *-holic* occurs after base words that end in a vowel.

there is no consensus on the morphological status of these group-forming elements in the literature, which is probably to a great degree due to the fact that they show many similarities to other morphemes.

There is, however, general consensus among scholars that combining forms are bound word-forming elements in English that occur in word-initial or in word-final position. Hence, they are more accurately described by the terms 'initial combining form' (ICF) and 'final combining form' (FCF) respectively (cf. Bauer 1983: 214). The term 'combining form' was traditionally used to refer to the first or final elements in neo-classical compounds. These are compounds which have been "derived from a neo-latin or a neo-greek root" (Fischer 1998: 55). Neo-classical elements are abundant in the English language, owing to its long history of borrowing.

However, linguists such as Beatrice Warren (1990), Adrienne Lehrer (1998) and others use the term to refer to elements of native English origin as well as to neo-classical morphemes. Filtering the entries in the Oxford English Dictionary (OED) by selecting the part-of-speech category 'combining forms' in the search interface produces a list of 2259 results, the majority of which are of Latinate and Ancient Greek origin. *Aero-*, *bio-*, *geo-*, *-ology*, *-graphy* and *-morph* are only a few examples of well-established neo-classical combining forms in English. Yet there are also elements among the OED entries which are clearly derived from English words, such as *acousto-*, *Japano-*, *must-*, *-pager*, *-nap* and *-tainment*, amongst others. Merriam-Webster also uses this category to describe elements such as *Brit-*, *e-*, *-bot*, *-gate* and *-drome*, which are of native origin. As they share a large number of characteristics with their neo-classical counterparts, a definition of combining forms should include these native elements.

One of the major problems of most definitions is that combining forms share characteristics with other morphological elements from which they have to be delineated, namely affixes, clippings, roots, stems, and blends (see Fischer 1998 and Kastovsky 2009 for a more in-depth discussion of the delineation of combining forms from other elements). They are usually distinguished from affixes, clippings, roots, and stems by their ability to combine with affixes and with one another, their boundness, their lexical meaning and their ability to form compound-like formations, so-called combinations (cf. Fischer 1998; Prćić 2005). A blend may be the source of a new combining form if the splinters present in the blend, i.e. the incomplete parts of the source words (see Bauer 2006), become productive. However, blends cannot be said to consist of combining forms as they usually do not give rise to productive word-formation patterns and the splinters in the blend generally do not occur in other words. Unless the blend formatives are used productively to coin analogous words, i.e. words containing the same element such as *bagelicious* and *turkeylicious* in analogy to *delicious*, they cannot be considered combining forms (Warren 1990).

Despite the fuzzy definitions and consequent multifaceted nature of the elements subsumed under them, combining forms exhibit a number of distinctive characteristics (see Fischer 1998: 55–57 and Lehrer 1998: 14, whose descriptions form the basis for the following list and who may be referred to for a more detailed account). The following are typical characteristics of combining forms:

1. Combining forms are necessarily bound. Unlike clippings, they cannot stand alone.[6]
2. They occur word-initially (initial combining forms) or word-finally (final combining forms). Some neo-classical combining forms may occur both word-initially and word-finally (cf. *morphology* and *biomorph*).
3. They may combine with both free and bound elements, which can be
   a) other combining forms (e.g. in *geography*, *telethon*);
   b) free lexemes (e.g. in *astrophysics*, *Frankenfruit*), including clippings;
   c) affixes (e.g. in *a-morph-ous*).
4. They occur in compound-like formations called combinations (cf. Fischer 1998).
5. They may be either:
   a) borrowings from the classical languages adapted to English (neo-classical combining forms); or
   b) native elements formed from English source words.
6. Initial combining forms often end in a vowel, usually *-o*. If the medial vowel is morphologically triggered or not assignable to either element, it is a linking vowel.
7. They carry lexical, not grammatical, meaning[7].
8. They are word-forming elements which are currently productive or were productive at some point in the past.

[6] Some combining forms undergo a process of conversion in which distinct free lexemes of the same meaning are created, e.g. *burger* from *-burger*. These free lexemes can then no longer be considered combining forms (see Fischer 1998: 57).

[7] This criterion poses problems insofar as ascribing lexical meaning implies that there is a measure of lexical density, but no criteria for this have yet been defined. Furthermore, there are affixes which also carry lexical meaning, cf. *super-*. Nevertheless, the criterion will be retained as it serves to distinguish combining forms, which carry lexical meaning comparable to nouns and adjectives, from affixes, which carry functional meaning comparable to prepositions, adverbs or numerals.

These characteristics can be used in lieu of a definition in order to discriminate them from other morphological elements.

3 Combining forms and their productivity

Native combining forms as a morphological category are reasonably well documented lexicographically. Recent research in this field mostly focusses on their demarcation from other morphological categories such as affixes and blends (cf. Fradin 2000; Kastovsky 2009; Prćić 2005, 2008). The acceptance of combining forms as a category of morphological elements is further corroborated by the fact that 'combining form' is used as a morphological category in dictionaries; for example, it is one part-of-speech tag used in the OED next to 'prefix' and 'suffix'.

These elements, "which linguists intuitively feel are neither affixes nor roots" (Warren 1990: 115), have been examined in a number of lexicographic studies. Warren (1990) proposes five subcategories of combining forms and discusses the processes of secretion and abbreviation from which they may result, followed by a discussion of their morphological status. Back (1991) provides a detailed description of neo-classical, pseudo-classical, and native combining forms, which he subdivides into simplex and complex elements and refers to as 'non-neoclassical', and their respective properties such as structure, distribution, and semantics. There are also surveys of individual native combining forms such as *-(a)holic* (cf. Algeo 1981; Kolin 1979). The semantics of native combining forms are investigated by Lehrer (1998), who uses the term 'combining form' to refer to the parts of the source words that form a blend, e.g. the splinters *sm-* and *-og* in *smog*. In her view, however, a blend formative can only be considered a combining form if it becomes productive "and since productivity is a matter of degree, there is a scale from highly productive morphemes like *-holic* to splinters that have been used only once (apparently), like *-nography* in *warnography* < *war* + *pornography*" (Lehrer 1998: 4–5). There is no discussion of what may trigger productivity in blend formatives, but Lehrer speculates that productivity in splinters is dependent on "nonlinguistic, mostly chance factors" (1998: 5). Her analysis of the semantics of these elements suggests that high salience in the source word may be a factor in their productivity, as this helps speakers identify the source word in the resulting neologism and decode the meaning of the new formation (cf. Lehrer 1998: 7). Unfortunately, there is no study to date relating the salience of combining forms as established by Lehrer to their actual productivity.

Despite the abundance of lexicographic surveys, there is very little empirical research on the productivity of the elements subsumed under the category 'combining forms'. Most research on productivity focusses on affixes (see, for example, Baayen & Lieber 1991). However, given the prominence of combining forms in the English language, such an investigation is certainly called for. A study of the productivity of the category of neo-classical combining forms is presented by Bauer (1998), who regards neo-classical compounds as "compromises" as they are "formed from foreign elements but as words of English" (Bauer 1998: 413). However, as the focus is on the productivity of the class rather than the elements subsumed under it, the study does not provide any insights into the current productive use of neo-classical combining forms.

The global productivity P* of native combining forms is investigated by Fischer (1998). The measure P* was proposed by Baayen and Lieber (1991) and takes into consideration the number of hapaxes per total number of types. Fischer bases her analysis of the morphological class on a sample of 100 combining forms collected from four dictionaries which were to be considered productive at the time of collection. Discussing their potential productivity, Fischer claims that if a dictionary entry exists for a combining form, its productivity can be assumed to be high in comparison to elements which are not listed (1998: 63). She explains, however, that some combining forms she considers productive such as *-flation*, *-crat*, and *-nomics* do not have their own dictionary entries, which she attributes to their being in a transitional phase between blend formative and combining form. Her subsequent study of the neologisms of *cyber-*, *techno-*, *info-*, *docu-*, *-umentary*, and *-tainment* (1998: 141–170) in *The Guardian* between 1990 and 1996 and *The Miami Herald* from 1992 onwards shows that *techno-* and *cyber-* are the most productive combining forms in the sample. In the case of *cyber-*, this productivity is explained by its topicality in the investigated period due to the importance of cybernetics and virtual reality across disciplines from the 1980s onwards (see also Coe 2015 for a discussion of the origin and rise of *cyber-*). Topicality is also found to be the reason for the higher productivity of *info-* as compared to *docu-*. The number of new formations varies remarkably in each year of the period under investigation as the elements may become fashionable within a short period of time: "[t]his sudden topicality can then have an effect on the number of new formation [sic!] created and result in large frequency fluctuations" (Fischer 1998: 163). Fischer relates the productivity of combining forms to their degree of institutionalisation. Institutionalised lexical items, according to a definition by Bauer (2003: 333), result from productive word-formation processes and are used regularly by members of a speech community, but have not yet become lexicalised. Generally, Fischer's study shows that highly institutionalised combining forms are highly productive. This applies, for example, to *techno-* and *cyber-*, which are widely and regularly used and whose meanings are assumed to be known and are thus not paraphrased by speakers (Fischer 1998: 152). The least productive combining forms in the sample, *-tainment* and *-umentary*, are also the least institutionalised.

Given that productivity is a time-dependent variable (Plag 2006), it is important to consider the use of an element over a longer period of time to draw conclusions as to its productivity. While Fischer's (1998) study provides an insight into the productivity and its relation to their institutionalisation of six selected combining forms during the early 1990s, she concedes that "a longer time span of examination is more fruitful, as the number of types can be expected to turn out higher within a longer than within a shorter period of time" (Fischer 1998: 166). Thus far, no comprehensive diachronic study on the productivity of native combining forms spanning the decades from their first attestation to the current time has been carried out. As Fischer's set is relatively small and the time span short, it is furthermore difficult to draw conclusions from her data as to the factors causing the diverging productivity of these elements, both diachronically and synchronically. The study presented in this chapter aims to fill this gap.

4 A corpus-based study on the productivity of native combining forms in American English

The present study investigates the productivity of native combining forms in American English. In a diachronic approach, the productivity of a sample of native combining forms is measured for the period from 1950 until 2009. The aim of the study was to determine whether the native combining forms investigated are currently productive and how this productivity changed over time. For this purpose, the neologisms in each decade were counted in COHA beginning with the one in which a given combining form was first attested. A further aim was to identify factors which influence the productivity of native combining forms and to explain why some are more productive than others.

The study was based on the following hypotheses:

1. Native combining forms are productive word-forming elements in American English.
2. The diachronic productivity of native combining forms depends on cultural factors. They are coined to name new phenomena, for example those brought

about by new technological developments and the new media. Consequently, if their referent is no longer of cultural importance, the productivity of the native combining form ceases.

4.1 Methodology

Drawing on the characteristics defined in Section 2, the sample of native combining forms was composed from the lists of combining forms by Warren (1990), Lehrer (1998), and the entries in the OED [8] (see Section 4.2). In order to allow for a diachronic study of the combining forms' productivity, the study sample was compiled of native combining forms which were first attested between the 1950s and the 1980s. The elements taken from the three lists were checked for the date of their first occurrence in COHA to determine whether they should be included in the sample.

The COHA was chosen because it is a very large corpus; it contained 400 million words at the time this study was conducted. It is composed of texts from magazines, newspapers, fiction, and non-fiction books from 1810 until 2009 and is the only up-to-date corpus with historical data. The texts contained in COCA; (Davies 2008–) are also included in the COHA. The COHA does not contain spoken text, which means that an important register of the English language cannot be taken into account. This is disadvantageous for research on combining forms because productivity is always dependent on register, as shown by Plag, Dalton-Puffer, and Baayen (1999). Unfortunately, there is no current and diachronic corpus of English that includes spoken data. As it only provides information on American English, the results of any research based on the COHA cannot be generalised to other varieties of English and so are somewhat limited in their significance. The British National Corpus (BYU-BNC), a possible alternative, has not

[8] As the characteristics of combining forms which are used for operationalisation here draw on the definitions by Warren (1990) and Lehrer (1998), there is considerable overlap between their lists and the one compiled for this study. Those elements of neo-classical origin and those which were coined before the 1950s or after the 1980s were excluded here, however. The OED was used in addition because it contains more recent coinages and employs a broader definition of the term 'combining form' than Warren and Lehrer, which means that more elements are listed for this category. This is advantageous for a comprehensive study of combining forms because it is difficult to compile a complete list of these elements, especially if they are infrequent, and the broad definition applied by the OED makes it more likely to contain elements which could qualify as combining forms as per the operational characterisation in Section 2.

been updated since 1993 and therefore cannot be used for determining the current productivity of word-forming elements. The Google Books Corpus (Davies 2011–) is considerably larger than the COHA and was also briefly considered for study, but while there are more hits for well-established formations, very few neologisms are found in this corpus. This is due to the fact that it contains neither newspapers nor magazines, in which the language is often more colloquial and therefore closer to spoken language. In fiction and non-fiction books, of which the Google Books Corpus is composed, authors are apparently more likely to use established forms and only rarely coin neologisms using the combining forms under investigation here. In some searches, no hapax legomena were found for a certain decade in the Google Books Corpus whereas the COHA produced more than a dozen, which suggests the latter holds more promise for this research project. Of course, every corpus is finite and there is always the possibility that certain elements do not occur at all despite speakers using them.

Once the sample had been compiled, each of the combining forms in the sample was entered into the COHA using a wildcard search (e.g. "heli*") limiting the number of hits to 10,000. This did not affect the results of this study, as none of the searches produced more than 2,000 hits. All of the entries that did not constitute types of the respective combining form were removed from the data, i.e. words which coincidentally begin or end in the same sequence of letters. The types were counted for each element to determine the type frequency. Plural forms were only counted as a type if the singular form did not occur in the list produced by the COHA. Only the entry with the earlier date of first attestation was counted as a type if both the singular and the plural were listed. Likewise, out of several spelling variants, only the earlier one was counted. The rationale for each of these decisions was that though their inflectional or orthographic properties may vary, such items constitute instances of the same type. Plural forms and spelling variants of elements occurring in the list were not discarded despite the fact that they do not constitute separate types. Instead, the entries were combined and their frequencies were added up in order to obtain a correct token frequency for the respective element. The overall token frequency was then calculated using the figures produced by the COHA.[9] As a rule, the source word was not included in the type frequency as it gave rise to the combining form, but is not a coinage itself. Once the type and token frequencies had been determined, the new formations occurring for the first time in the corpus were counted for each decade from the 1950s until the 2000s and then plotted. The frequency measures of the

[9] Each elements' type and token frequencies for the overall investigated period and for each decade are provided in the appendix.

elements were then compared to one another to establish which combining forms are diachronically productive and which are not.

This diachronic approach was chosen because in a synchronic approach, an element is usually considered productive if it has produced a high number of types, especially hapax legomena. However, it might be the case that almost all of these types first occurred thirty years ago and no new formations have been created since. In such a case, the total number of types and tokens in the corpus would be misleading and assessing such an element as currently productive would therefore be a false conclusion. For this reason, the general productivity P* is not used here. A diachronic survey of the number of neologisms per decade is a more adequate measure to determine the current productivity and to evaluate how this productivity has changed in the past and might develop in the future. Finally, based on the evaluation of the productivity of the combining forms in the sample, the productive elements were compared to the unproductive ones to establish semantic and pragmatic factors which influence productivity.

4.2 The study sample

The study sample contained 21 native combining forms (see glosses in Table 1).

Tab. 1: Study sample of combining forms first attested in the 1950s, 1960s, 1970s and 1980s.

First Attestation	Combining Form	Gloss	Source Word
1950s	heli-	'related to or relating to a helicopter; aircraft resembling helicopters'	helicopter
	-lect	'variety of the language'	dialect
	petro-	'relating to petroleum'	petroleum
	robo-	'robotic or automatic X'	robot
	-think	'characteristic mode of thinking'	doublethink
1960s	cyber-	'related to virtual reality'	cybernetics
	eco-	'ecological(ly)'	ecological
	-holic	'addicted to X'	alcoholic
	hover-	'utilising an air-cushion as a means of support'	hovercraft

First Attestation	Combining Form	Gloss	Source Word
	-jacking	'theft, seizure'	hijacking
	-pager	'having a specified number of pages'	–
	-speak	'characteristic mode of speaking'	newspeak
1970s	docu-	'documentary'	documentary
	-gate	'scandal'	Watergate
	-jack	'to take illegally, to steal, to seize and commandeer, sometimes under threat'	hijack
	porta-	'portable'	portable
1980s	-babble	'confusing or pretentious jargon'	babble
	-bot	'robot or automated device'	robot
	-friendly	'adapted, suitable, useful or safe to X'	friendly
	info-	'information'	information
	-ware	'type of software'	software

The combining forms surveyed here were first used as bound lexical elements between the 1950s and 1980s. These decades were chosen because enough time has passed since to allow for an evaluation of the diachronic development of the productivity of combining forms. Furthermore, only a handful of native combining forms were first attested before the 1950s. The elements were collected from Warren's (1990) and Lehrer's (1998) papers as well as from the OED. Each of the elements listed in these sources was entered into the Corpus of Historical American English (COHA) in order to determine the decade in which it first occurred as a bound lexical element in a formation other than the source word. Those which first appeared in the 1950s, 1960s, 1970s or 1980s were included in the study sample[10]. Many of these elements occur in all three sources and have their own dictionary entries, although some are listed as affixes (e.g. *-teria* and *-(a)holic*).

[10] There are combining forms whose first occurrence in the corpus was earlier than the 1950s or later than the 1980s and which are not included for that reason, for example *bio-*, *-thon* and *-scape* on the one hand and *-licious* and *Franken-* on the other. This is not to say that they are not currently productive. Moreover, earlier quotations are given in the OED for a number of the elements used in this study, which means that their first attestation diverges from the one quoted here. For a unitary approach, the decade of first occurrence in the COHA was taken as a point of reference.

As only native elements were to be examined, all neo-classical elements, that is those that have a counterpart in the classical languages, were excluded from the sample. Furthermore, scientific and technical terms were not included. They are predominantly of Latin or Ancient Greek origin and in many cases dictated by institutions. One element, namely *-schooler*, was not included in the sample even though it first occurred in the corpus in the 1970s and is labelled in the OED as a combining form. It is not regarded as a combining form here because all the words in which it occurs (*high-schooler*, *pre-schooler*, *grade-schooler*, *home-schooler* etc.) can be regarded as derivatives which have been formed by the addition of the derivational suffix *-er* to the respective compounds of *school*.

A number of the combining forms cited in the OED did not produce any results in the COHA and were consequently disregarded in the sample. Their absence from the corpus may be coincidental and ascribable to the finiteness of the COHA, but it is quite likely that the elements which could not be found in the corpus are in fact unproductive. Two such examples are *acousto-* and *syntacto-*.

5 Results

The combining forms in the study sample are as heterogeneous in the number and kind of formations they give rise to as in their structure. The type frequency varies from three combinations coined with *-jack* to 178 combinations coined with *cyber-*. The number of formations per element[11] are shown in Figure 1 below:

[11] The COHA data for each combining form in the sample is provided in the appendix.

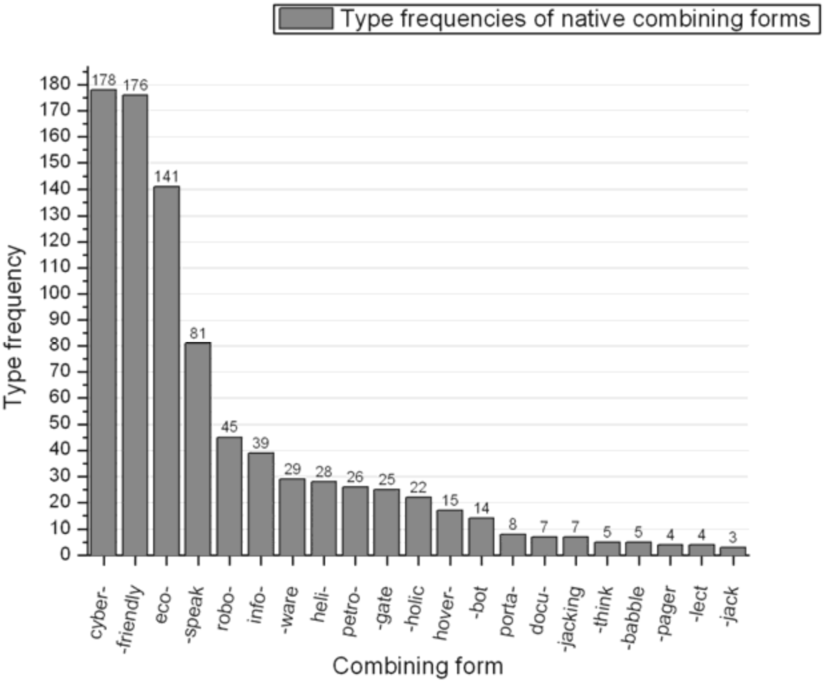


Fig. 1: Type frequencies of native combining forms in the study sample.

The results of the general frequency query attest that the elements in the sample differ greatly with regard to their occurrence in the corpus, which is a first indication of their diverging productivity. To provide a comparable overview of the respective numbers of neologisms, the new formations each element has produced per decade were plotted. The following 3D bar charts (Figures 2, 3, 4, and 5) show the new formations each element has produced per decade. For reasons of lucidity, the combining forms are grouped by the decade in which they were first attested in the COHA. In all charts, the highest increment on the Y-axis is 105. It was chosen because the highest number of neologisms observed in one decade in the study sample was 101 (*-friendly* combinations in the 2000s). The number of formations coined using the other combining forms in the sample can be compared to this number.

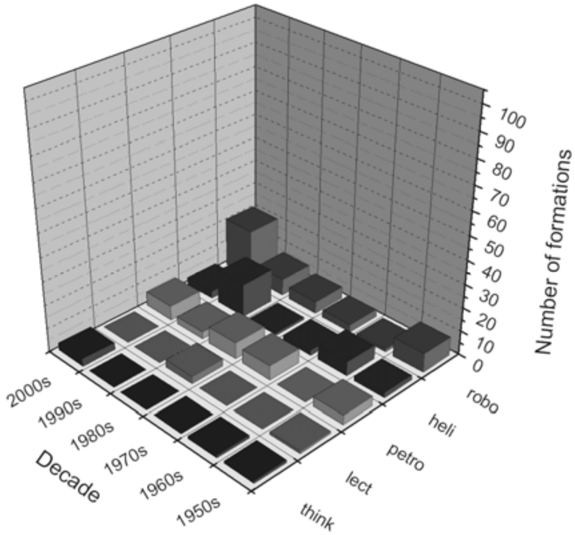


Fig. 2: Combining forms first attested in the 1950s: new formations by decade.

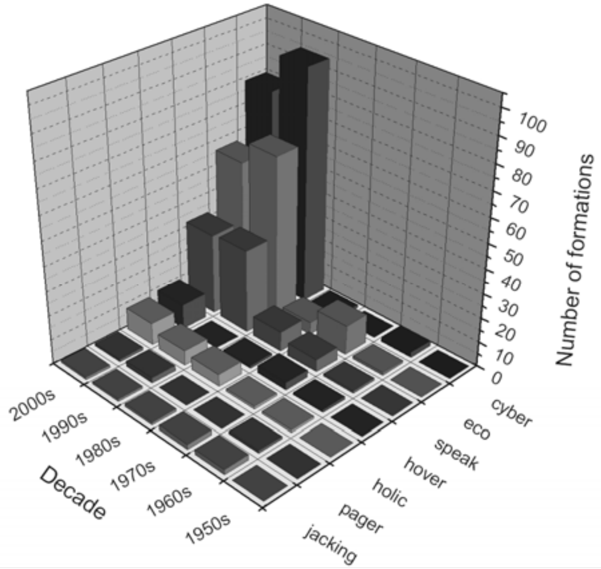


Fig. 3: Combining forms first attested in the 1960s: new formations by decade.

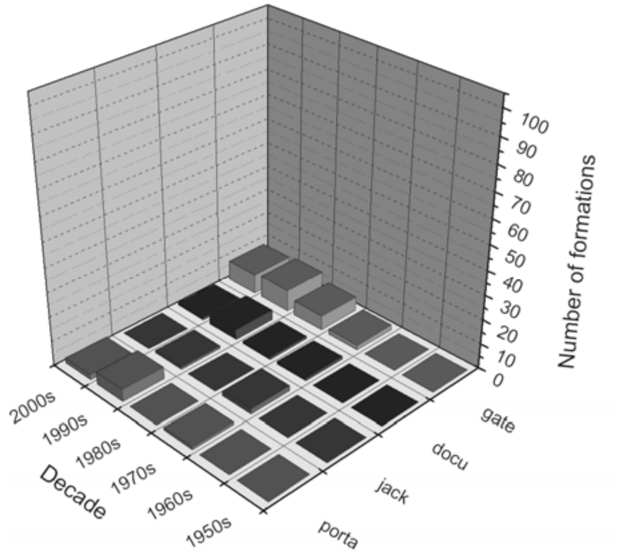


Fig. 4: Combining forms first attested in the 1970s: new formations by decade.

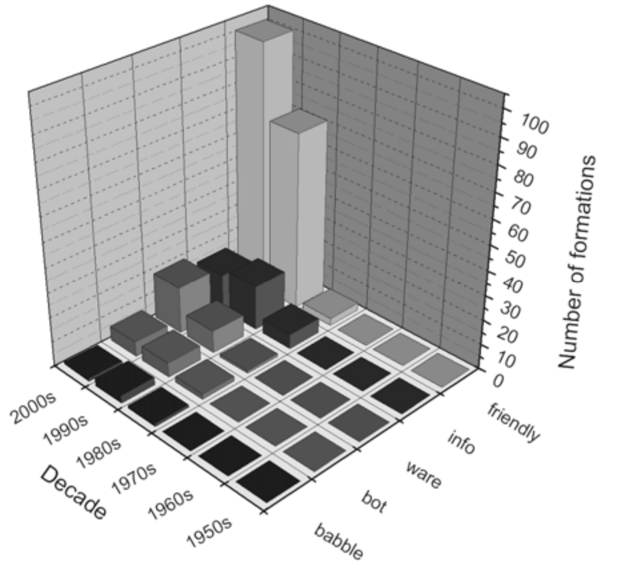


Fig. 5: Combining forms first attested in the 1980s: new formations by decade.

Some general conclusions can be deduced from these graphs. Firstly, the number of new formations per decade has increased over time for most elements. Consequently, native combining forms are generally not a temporary fashion which only lasts for a few years. Many of them are apparently here to stay. Secondly, eleven elements produced their highest number of neologisms in the 1990s. This indicates that in general, the 1990s were a decade in which combining forms were frequently used. This might be due to the technological progress and the rise of the Internet in that decade. Six native combining forms produced the largest number of neologisms in the 2000s, which shows that they are still very popular. Apparently, native combining forms in general gained popularity from the 1970s onwards. Thirdly, eight combining forms, not an insignificant number, have produced only ten or less neologisms in the period investigated. Two out of these eight, namely *-lect* and *-jack*, have not produced any new formations in the past decade. The combining forms *-friendly*, *cyber-*, and *eco-* display the highest total number of types. The element *-friendly* has also produced the highest number of neologisms in one decade: 101 new formations in the 2000s.

Considering the diachronic change in the production of new forms for the elements and their total type frequencies (compare Figure 1), different patterns of the appearance of neologisms are observed. The combining forms can be grouped according to these patterns.

The first group of elements leads to the formation of a small number of combinations per decade but their total type frequency is low. This pattern can be observed in the combining forms *-bot*, *-babble*, *docu-*, *hover-*, *-jack*, *-lect*, *-jacking*, *-pager*, *porta-*, and *-think*. The number of types each of these elements produced over the decades ranges between three and fifteen. Some do not produce any new forms in some decades but then reappear. The combining forms -*lect* and *-jack* are extreme cases. They only produced four and three types respectively and no new formations have been recorded since the 1990s.

It is quite remarkable that the element *-bot* has made its way into the English language considering that there is already a combining form abstracted from the same source word and of the same meaning, namely *robo-*. It appears as if *-bot* formations are mainly used to refer to automated technical tools and devices which serve a specific purpose (e.g. *spybot*, *drill-bot*, *packbot*), whereas *robo-* coinages appear mostly in fiction novels and predominantly describe robotic creatures (e.g. *robocop*, *robo-warrior, robo-dog*, *robo-parrot*). Their use in different registers could explain the coexistence of the two apparently synonymous combining forms. A further difference is that the initial combining form *robo-* does not combine with adjectives but the final combining form *-bot* does, for example in *hotbot*.

The second group steadily produces new forms but there is no substantial rise in the number of neologisms per decade over time. Nevertheless, the total number of types is considerably higher than that of the first group. They have now produced twenty to fifty words. The combining forms *-gate*, *heli-*, *-holic*, *info-*, *petro-*, *robo-*, and *-ware* display this pattern. In one decade, there might be a considerable number of neologisms, but this could be followed by a decade of fewer or even no new forms. If a curve were fitted, it would rise gradually or remain nearly stationary. Apparently, some of the elements have gained popularity in the past two decades. Formations with *-ware* such as *student-ware*, *cinemaware*, and *spyware* seem to be following an upward trend and it is possible that an even larger number of neologisms will be coined in the current decade.

Finally, there is a small group of elements which initially only produced a few new words. The number of attested neologisms then suddenly rose steeply from one decade to another. These elements are *cyber-*, *eco-*, *-friendly*, and *-speak*. Their total type frequencies exceed 80, which is comparatively high. *Cyber-* and *-friendly*, as mentioned before, have given rise to more than 170 words, which is extraordinary for any combining form. Amongst them are established formations such as *cyberspace* and *user-friendly*, but also more unexpected types such as *cybervandalism* and *ozone-layer-friendly*. The polysemic combining form *-ware* might be on the brink to entering this third group. It is a relatively new combining form used in formations such as *spyware* and *freeware* and there is constant progress in the fields of software and merchandise, so more *-ware* neologisms might be coined in the next years.

It is noteworthy that *-speak* has produced considerably more formations than *-think* even though they sprung from the same literary source, George Orwell's *Nineteen Eighty-Four*, and refer to somewhat related concepts. This may be due to the fact that the verb *think* already has corresponding nouns in *thought* and *thinking*, both of which are also related in morphological make-up. The corresponding noun to *-speak*, on the other hand, is obsolete (compare the OED entry for *speak*, n.). The noun *speech* has a different meaning; it does not refer to a way of speaking, and *jargon* is much more specialised and potentially contemptuous. Consequently, the combining form *-speak* may have filled a void in the language. This would explain the comparatively large number of neologisms, e.g. *FBI-speak*, *geekspeak*, and *winespeak*.

In summary, the data has shown that all of the elements have produced new forms in the period examined, especially in the 1990s and the 2000s, but the numbers vary greatly. How the type and token frequencies recorded relate to productivity and how the productivity of certain native combining forms can be explained will be discussed in the following section.

6 Discussion

In the following sections, the results presented above will be discussed in the light of the hypotheses as to productivity of native combining forms that were at the core of this study. Section 6.1 addresses the question of whether the elements in the set can be considered productive morphemes. Section 6.2 is devoted to an examination of semantic and pragmatic aspects that either boost or restrict the productivity of native combining forms.

6.1 Are native combining forms productive morphemes?

The question of whether native combining forms are productive cannot be answered in general. Evaluating the data collected from the COHA shows that the number of new formations created using the investigated native combining forms differs enormously from one element to the next and even for each element from one decade to the following.

The number of neologisms coined using a combining form is indicative of the productivity of this element in the respective decade. Consequently, an element which produces no new forms in a specified time span is not productive in that period. Those combining forms that have produced neologisms can be said to be productive to a certain degree: they may display a low, a moderate or a high productivity. These labels can only be applied in comparison with other combining forms and only specify certain points on a scale between ‘unproductive’ and ‘fully productive’. The description of a combining form’s productivity based on the number of neologisms per decade is only a relative measure. There is, of course, a possibility that the findings are distorted by the types of texts in the COHA, so that the conclusions drawn here can only be generalised for the text types found in this corpus.

Based on the development of the type frequency per decade, the study sample can be divided into three groups of elements. Each of the groups displays a different pattern with regard to the frequency of new formations appearing.

The combining forms *-bot*, *-babble*, *docu-*, *hover-*, *-jack*, *-jacking*, *-lect*, *-pager*, *porta-*, and *-think*, the elements from the first group, have to be regarded as having only low productivity. They only give rise to a small number of six or less neologisms per decade. Their overall type frequency is 15 or lower. Based on the corpus data from the year 2000 to 2009, the combining forms *-lect* and *-jack* should currently be considered unproductive. No new forms were attested in that period. They might produce neologisms in the future, but their productivity may also have ceased.

The second group of elements contains *-gate*, *heli-*, *-holic*, *info-*, *petro-*, *robo-*, and *-ware*. These combining forms are moderately productive. Their overall type frequency is above 25 and they have produced a more or less steady output of neologisms over the period investigated. Unlike the third group's, these elements' type frequencies have not risen substantially over the decades in the investigated period.

The final group consists of the elements *cyber-*, *eco-*, *-gate*, and *-speak*. These are highly productive combining forms that have all produced at least eighty types from the time they first appeared until the 2000s. Their token frequencies exceed 100 – or, in the case of *eco-*, even 1,000. The number of neologisms for each of these combining forms has risen enormously from one decade to the next in the period examined. It can be expected that the curve will reach a plateau or fall again at some point. This might already be in progress for *cyber-* and *eco-*. Nevertheless these elements can be regarded as highly productive – at least for the time being.

The investigation has shown that as a category of morphological elements, native combining forms are productive. They are used in the analogous coinage of neologisms. It has also been documented that the processes of creating native combining forms are productive and speakers tend to invent new ones. At least four new combining forms have appeared during each of the decades examined and the list is far from complete. It is likely that there are more native combining forms that became productive during the period examined; however, unless they appeared in the source from which the study was compiled, they were not considered.

This investigation confirms the first underlying hypothesis. All of the elements investigated have produced at least three types, some more than 100, from 1950 to 2009. Consequently, all combining forms in the study sample have been productive at some point in the surveyed period. Since they are or have been productive, they are indeed combining forms as per the definition. However, not all of the combining forms examined are currently productive, as some elements examined did not produce any neologisms in the past decade.

6.2 Explaining the productivity of native combining forms

A combining form cannot arbitrarily be attached to any given word to coin neologisms: the coinage of new words is semantically and pragmatically restricted (see Plag 2006). Neologisms are coined if they are useful for speakers, e.g. because they denote something for which there previously was no word. A further motivating factor may be speakers' desire to be extravagant, i.e. to make their utterances interesting so that they are noticed for their creativity by others (see Haspelmath 1999 for a more detailed description of extravagance as a motivating factor in neology). Fashionable and innovative terms are likely to catch on and lead to analogous formations by other speakers, especially if they are perceived as expressive. As a result, fashion and extravagance can positively influence productivity. On the other hand, new formations may be avoided by speakers because they do not have a referent or because a synonym or a homonym already exists in the language. For example, *thief* generally blocks the possible synonym **stealer*. The existence of a homonym makes it difficult for members of a speech community to decode the meaning of a new formation unambiguously, which is potentially detrimental to its productivity. Moreover, a combining form may not combine with certain words because there are structural restrictions imposed on the bases to which it may attach. These may be phonological, morphological, syntactic or semantic in nature (cf. Plag 2003: 59–68). The more restrictions there are on the possible bases of a combining form, the less likely are neologisms and the lower is the productivity of the said combining form. As discussed above, not all native combining forms investigated here are equally productive. In fact, the number of neologisms produced by each element has been shown to vary greatly between high-productivity and low-productivity elements.

Several semantic and pragmatic factors influencing the productivity of these elements are conceivable. As mentioned before, a new formation has to fulfil a specific purpose for the speakers of the language. This is the case, for example, if it denotes something which had not previously been named (see Lehrer 2003). A formation is not useful if there is already a well-established element of the same meaning and function in the language. In this case, the doublet is blocked. This probably explains the lack of productivity of the elements *-jack* and *-jacking*. Their meanings are covered by the simplex words *steal* and *theft*. Both of these are very frequent in English and occur in compounds, so that there is no need for additional lexical items and *-jack* and *-jacking* are avoided. In some cases, *-jacking* and *-jack* formations refer to incidents of theft under threat, e.g. *carjacking*, *shipjack*. The relative scarcity of events of that nature also explain the low productivity of these two elements. The combining form *-bot* may even be blocked by its source word *robot* because it is semantically transparent and has only one extra syllable. These three combining forms probably display such a low productivity because speakers do not need them. Combining forms, like affixes, are of course limited with regard to the number of possible bases to which they attach.

If the meaning of a combining form is very narrow, it is unlikely that it will produce many formations and its productivity is expected to be low. Homonyms in the language can also have a negative effect on the productivity of the combining form. The low productivity of *-jack* and *-pager* may be in part due to the fact that there are homonymous words which are much more frequent in English. Similarly, the letter sequence *-lect* (*idiolect*, *basilect*) also occurs in other words (*intellect*, *neglect*, *elect*) which veils its meaning and makes it less salient.

However, in some cases a (near-)synonymous form does become a part of the language's vocabulary. This can happen because speakers do not know or cannot activate the rival word. If the new formation is shorter than the existing word, follows regular grammatical rules or has a different function, speakers may prefer it to the established alternative. The productivity of abbreviated combining forms such as *eco-* and *info-* can be thus explained. Their creation and use is apparently a matter of language economy: using short and catchy combining forms (*ecobricks*) is more economical than forming compounds of the same meaning from the source words (*ecological bricks*). Polysemic combining forms differentiate themselves from their free counterparts by their boundness and altered meanings. They are useful and can thus become established.

When considering the semantic features of the native combining forms, it is striking that many have meanings relating to the Internet and information technology. Countless combinations from native combining forms occur in online media. Apparently, speakers also have a tendency to use combining forms to refer to phenomena from this field. The elements *cyber-*, *-bot*, *info-*, *robo-*, and *-ware* are semantically connected to the fields of virtual reality and artificial intelligence (cf. *cyberworld*, *anthrobot*, *info-war*, *robo-pets*, *virtuware*). All of them are relatively productive.

In general, the productivity of any given morphological element is subject to pragmatic factors (cf. Plag 2006). One requirement is that they denote something nameable. The productivity of a combining form is dependent on its meaning especially in those cases where the meanings relate to current phenomena, for example those brought about by progress in technology or the sciences, or topics of public discourse. As a consequence, their productivity is closely linked to the currentness of the phenomenon they denote. Usage of a given morpheme can be influenced by fashion and may fluctuate. Consequently, the productivity of combining forms is also a matter of fashion. Some combining forms quickly disappear when their denotatum is no longer of cultural interest. This affirms the second underlying hypothesis of this study and can be observed in the combining form *porta-*. When first introduced to the markets in the 1960s, portable gadgets and appliances were a novelty, cf. *portacrib*, *portaphone*. Companies made use of the combining form to coin catchy names (cf. *porta-potty*). Nowadays, portability is no longer regarded as unusual but has become the norm for many devices. This explains the low productivity and infrequent use of this combining form in recent years. Reversely, the high productivity of the combining forms *cyber-*, *eco-* and *-gate* can also be assigned to cultural factors. Globalisation and the development of the Web 2.0 and social media mean that a large part of our lives now takes place online. Consequently, there are many new phenomena which can be aptly named using *cyber-* combinations, e.g. *cybersex*, *cyberbullying*, and *cyberstalking*. Similarly, awareness of the effects of pollution and waste on the environment has generated a need for terms relating to environment-friendly politics, behaviour and materials which was met by *eco-*, e.g. in *eco-awareness*, *eco-adventures*, and *eco-timber*. The emergence of *-gate* as a productive combining form was rendered possible because it became fashionable to use *-gate* terms after the Watergate scandal. A new combining form may be coined as a side effect of a current issue or phenomenon and disappear as soon as the phenomenon subsides (e.g. *doughnutgate*), or it may become established so that speakers continue to use it (e.g. *Monicagate*).

In summary, semantic and pragmatic factors clearly affect the productivity of native combining forms. Their meaning is an important factor in this. Those elements referring to information technology have been shown to have a high productivity. If there is already a synonym or a homonym in the language, a morpheme may be avoided which decreases its productivity. Combining forms are subject to fashion trends. They often have meanings related to current issues. If they refer to temporary phenomena, they can quickly become unfashionable. Narrow meanings also limit the productivity of an element.

7 Conclusion

Combining forms are a productive class of bound word-forming elements in English. They are versatile morphemes which are readily used to form creative neologisms. They occur in a large variety of contexts and are often employed to catch the eye of the reader through their unusual form.

This corpus-based study on the productivity of native combining forms based on COHA data has revealed that, generally, combining forms can be considered productive in word-formation. All of the elements investigated displayed productivity at some point in the period from 1950 until 2009. This confirms the first hypothesis on which the study was based. However, while some combining forms have produced more than a hundred new words since they were first created, other elements have only produced a handful and some are not currently productive. The investigation has shown that this can be attributed to semantic and pragmatic factors.

The semantic fields in which a combining form is used determines its productivity. Many combining forms relate to the fields of information technology and virtual reality and are frequently used on the Internet. The semantics of a combining form can also have a negative effect on its productivity. This is the case when there are more frequently used synonyms or homonyms already established in the language which obscure the meaning of the combining form and make it difficult for speakers to decode it. This can have the consequence that the combining form is blocked, leading to low productivity.

Combining forms are often used in eye-catching terms relating to current phenomena and issues, predominantly in the media and online. Their productivity and frequency are closely connected to the topicality of the denoted concept and are subject to fashion. In this aspect they are similar to other creative neologisms such as acronyms, blends, clippings, and lexical phrases (cf. Fischer 1998). The productivity of native combining forms is indeed in some cases determined by cultural factors. For example, in the 1990s there was a remarkable rise in the number of combinations referring to Internet-related phenomena. The productivity of some combining forms decreases as soon as their referent goes out of fashion. However, they usually do not disappear immediately and it is still possible that neologisms using a particular combining form are coined in the following decades. There are also combining forms that enter the language as fashionable and creative coinages but become established and retain a moderate or even high productivity.

The COHA has proven to be a useful corpus for this study. It contains a large number of words from various text types so that the results can be considered rather representative at least for the written registers of American English. The diachronic structure of the corpus made it possible to compare different decades and the respective frequencies which was shown to be insightful and revealed that the productivity of combining forms may fluctuate.

The results of the study can only be generalised to American English and only to the types of texts contained in the COHA. As combining forms are abundant in social media, which were not considered in this investigation, this study may provide an incomplete and possibly misleading picture of the actual productivity of combining forms. It is probable that the numbers of neologisms found on the Internet are much higher than those in the consulted corpus. Additionally, there are more native combining forms in English which were not considered here. Investigations on the productivity of native combining forms in the current decade, on the Internet, in spoken language, and in other varieties of English could be interesting future research projects.

Combining forms are increasingly important in English word-formation because they can be used to add lexical meaning to free lexemes that cannot be appropriately expressed by established affixes. They are likely to continue to shape the English language in the future as more creative neologisms appear and become lexicalised.

References

Adams, V. (1973). *An Introduction to Modern English Word-formation*. London, UK: Longman.

Algeo, J. (1981). More holics. *American Speech, 56*(2), 152–53.

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics, 29*(5), 801–844.

Back, G. (1991). *Combining Forms im Englischen*. Retrieved from https://web.archive.org/web/20120116161823/home.arcor.de/gernotback/Magister/Combining_Forms.html (last accessed October 2018).

Bauer, L. (1983). *English Word-formation*. Cambridge, UK: Cambridge University Press.

Bauer, L. (1998). Is there a class of neoclassical compounds, and if so is it productive? *Linguistics, 36*(3), 403–422.

Bauer, L. (2003). *Introducing Linguistic Morphology*. Edinburgh, UK: Edinburgh University Press.

Bauer, L. (2006). Splinters. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 77–78). Amsterdam, Netherlands: Elsevier.

Coe, T. (2015). Where is the origin of 'cyber'? *OxfordWords blog*. Retrieved from http://blog.oxforddictionaries.com/2015/03/cyborgs-cyberspace-csi-cyber/ (last accessed February 2017).

Fischer, R. (1998). *Lexical Change in Present-Day English. A Corpus-Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen, Germany: Gunter Narr.

Fradin, B. (2000). Combining forms, blends and related phenomena. In U. Doleschal & A. M. Thornton (Eds.), *Extragrammatical and Marginal Morphology* (pp. 11–55). Munich, Germany: LINCOM.

Haspelmath, M. (1999). Why is grammaticalization irreversible? *Linguistics, 37*(6), 1043–1068.

Kastovsky, D. (2009). *Astronaut*, *astrology*, *astrophysics*: About combining forms, classical compounds and affixoids. In R.W. McConchie, A. Honkapohja, & J. Tyrkkö (Eds.), *Selected Proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HEL-LEX)* (pp. 1–13). Somerville, MA: Cascadilla Proceedings Project. Retrieved from http://www.lingref.com/cpp/hel-lex/2008/paper2161.pdf (last accessed March 2017).

Kolin, P. C. (1979). The pseudo-suffix *-oholic*. *American Speech, 54*(1), 74–76.

Lehrer, A. (1998). Scapes, holics and thons: The semantics of English combining forms. *American Speech, 73*(1), 3–28.

Lehrer, A. (2003). Understanding trendy neologisms. *Rivista di Linguistica, 15*(2), 371–384.

Lehrer, A. (2007). Blendalicious. In J. Munat (Ed.), *Lexical Creativity, Texts and Contexts*. Amsterdam, Netherlands: John Benjamins.

Lüdeling, A. (2006). Neoclassical word-formation. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 580–583). Oxford, UK: Elsevier.

Plag, I. (2003). *Word-Formation in English*. Cambridge, UK: Cambridge University Press.

Plag, I. (2006). Productivity. In B. Aarts & A. McMahon (Eds.), *The Handbook of English Linguistics* (pp. 537–556). Malden, MA: Blackwell.

Plag, I., Dalton-Puffer, C., & Baayen, H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics, 3*(2), 209–228.

Prćić, T. (2005). Prefixes vs initial combining forms in English: A lexicographic perspective. *International Journal of Lexicography, 18*(3), 313–334.

Prćić, T. (2008). Suffixes vs final combining forms in English: A lexicographic perspective. *International Journal of Lexicography, 21*(1), 1–22.

Schmid, H.-J. (2005). *Englische Morphologie und Wortbildung. Eine Einführung*. Berlin, Germany: Erich Schmidt.

Warren, B. (1990). The importance of combining forms. In W. Dressler (Ed.), *Contemporary Morphology* (pp. 111–132). Berlin, Germany: de Gruyter.

Corpora and dictionaries

BYU-BNC	Davies, M. (2004–): *BYU-BNC. (Based on the British National Corpus from Oxford University Press)*. Available online at http://corpus.byu.edu/bnc/.
COCA	Davies, M. (2008–): *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.
COHA	Davies, M. (2010–): *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/.
Google Books Corpus	Davies, M. (2011–): *Google Books (American English) Corpus (155 billion words, 1810-2009).* Available online at http://googlebooks.byu.edu/.
Longman	Pearson Education Ltd. (2011): *Longman Dictionary of Contemporary English Online - LDOCE.* Retrieved from http://www.ldoceonline.com.
Macmillan Dictionary	Macmillan Publishers (2011): *Macmillan Dictionary and Thesaurus. Free English Dictionary Online.* Retrieved from http://www.macmillandictionary.com.
Merriam-Webster	Merriam Webster, Inc. (2011): *The Merriam-Webster Online Dictionary.* Retrieved from http://www.merriam-webster.com.
OED	Oxford University Press (2011): *The Oxford English Dictionary. OED Online.* Online versions July and September 2011 retrieved from http://www.oed.com.

Appendix: Data from the COHA

Tab. 2: 1950s combining forms.

Combining form		Total 1810–2009	1950s	1960s	1970s	1980s	1990s	2000s
HELI-	Token Frequency	218	2	42	10	19	93	52
	No. of new formations	28	2	6	2	1	14	3
-LECT	Token Frequency	8	1	0	0	7	0	0
	No. of new formations	4	1	0	0	3	0	0
PETRO-	Token Frequency	244	22	28	43	70	42	39
	No. of new formations	26	4	0	6	7	3	6
ROBO-	Token Frequency	96	26	5	5	9	17	34
	No. of new formations	45	8	1	3	5	7	21
-THINK	Token Frequency	26	1	2	1	3	4	15
	No. of new formations	5	1	1	0	0	0	3

Tab. 3: 1960s combining forms.

Combining form		Total 1810–2009	1950s	1960s	1970s	1980s	1990s	2000s
CYBER-	Token Frequency	672	0	8	2	1	334	282
	No. of new formations	178	0	2	0	1	95	80
ECO-	Token Frequency	1,155	0	2	56	64	409	624
	No. of new formations	141	0	1	12	5	66	57
-HOLIC	Token Frequency	129	0	1	2	28	27	71
	No. of new formations	22	0	1	1	5	6	9

Combining form		Total 1810–2009	1950s	1960s	1970s	1980s	1990s	2000s
HOVER-	Token Frequency	100	0	2	10	2	2	84
	No. of new formations	15	0	1	3	1	0	10
-JACKING	Token Frequency	45	0	2	15	5	6	17
	No. of new formations	7	0	2	2	1	1	1
-PAGER	Token Frequency	9	0	1	0	1	4	3
	No. of new formations	4	0	1	0	0	2	1
-SPEAK	Token Frequency	111	0	3	5	9	43	51
	No. of new formations	81	0	1	5	8	34	33

Tab. 4: 1970s combining forms.

Combining form		Total 1810–2009	1950s	1960s	1970s	1980s	1990s	2000s
DOCU-	Token Frequency	37	0	0	2	19	10	6
	No. of new formations	7	0	0	1	1	4	1
-GATE	Token Frequency	63	0	0	10	13	25	15
	No. of new formations	25	0	0	2	6	9	8
-JACK	Token Frequency	7	0	0	2	2	1	2
	No. of new formations	3	0	0	2	0	1	0
PORTA-	Token Frequency	41	0	0	1	0	34	6
	No. of new formations	8	0	0	1	0	5	2

Tab. 5: 1980s combining forms.

Combining form		Total 1810–2009	1950s	1960s	1970s	1980s	1990s	2000s
-BABBLE	Token Frequency	18	0	0	0	3	11	4
	No. of new formations	5	0	0	0	1	3	1
-BOT	Token Frequency	40	0	0	0	2	17	21
	No. of new formations	14	0	0	0	2	6	6
-FRIENDLY	Token Frequency	492	0	0	0	9	145	338
	No. of new formations	176	0	0	0	3	72	101
INFO-	Token Frequency	210	0	0	0	14	130	66
	No. of new formations	39	0	0	0	6	18	15
-WARE	Token Frequency	249	0	0	0	2	38	206
	No. of new formations	29	0	0	0	1	9	18

Mark Kaunisto and Juhani Rudanko

Advise against -ing: An emerging class of exceptions to Bach's Generalization

Abstract: This chapter discusses the occurrences of covert noun phrase objects with the verb *advise*. Contrary to the condition known as Bach's Generalization, according to which noun phrase objects cannot be omitted in object control structures, linguists have also noted that covert objects in structures of this type are indeed possible. With the aid of large electronic corpora, instances of the covert object pattern with *advise* are examined from the 19th century to the present day. In addition to quantitative observations on the use of the covert object pattern, attention is given to the semantic characteristics that the pattern manifests compared to instances with explicit objects. Pragmatic considerations may also be relevant in that the indeterminate specificity of the understood object could make the covert object pattern particularly suitable or desirable in situations where considerations of tact could play a role.

1 Introduction

The variety seen in the complementation patterns selected by different verbs in English is a subject that has received a great deal of attention from linguists in recent years. Verbs show different patterns of behaviour not only as regards the types of possible complements that they select, but also the numbers of arguments that they take. As will be observed in the present chapter, these characteristics themselves can change over time, and individual verbs may be in a state of flux in this regard. This chapter focuses on the verb *advise* and the possibility of leaving out the object when it is followed by a sentential *against -ing* complement. To take a closer look at the basic structure of the verb with this sentential complement pattern, we can first begin by considering sentence (1) from the Corpus of Historical American English (COHA; M. Davies 2010–):

(1) I strongly advised her against undertaking this play, [...] (COHA, 1949, FIC)

Mark Kaunisto, University of Tampere, mark.kaunisto@uta.fi
Juhani Rudanko, University of Tampere, juhani.rudanko@uta.fi

https://doi.org/10.1515/9783110489071-010

In example (1) the matrix verb *advise* selects three arguments. The first is the higher subject, realized by the NP *I*, the second is the direct object, realized by the NP *her,* and the third is the more complex complement *against undertaking this play*. The *-ing* form of the third argument is a gerund, and it is assumed here that the third argument involves an understood subject and is sentential in a sentence of the type of (1). The concept of an understood subject was made use of by traditional grammarians, including Jespersen, and it is made use of by many scholars today. It is also utilized by Huddleston and Pullum (2002: 65). A simple argument in favour of the assumption is the consideration that the lower verb needs a subject argument of its own. As Jespersen put it in 1940:

> Very often a gerund stands alone without any subject, but as in other nexuses (nexus substantives, infinitives, etc.) the connexion of a subject with the verbal idea is always implied. (Jespersen [1940]1961: 140)

For its part, the higher object of (1) receives a theta role from *advise*, which means that the construction of (1) is a control structure, rather than an NP movement structure (W. D. Davies & Dubinsky 2004: 3–4). In line with current work, the lower subject of (1) may therefore be represented by the symbol PRO. (For the framework used, see e.g. Carnie 2007: 395–6; Chomsky 1984: 20–21, Chomsky 1986: 118–130, and W. D. Davies & Dubinsky 2004). In example (1) PRO is controlled by the NP *her* in the higher clause, and the structure is one of object control. The constituent structure of example (1) may then be represented as in (1´) for the purposes of the current investigation:

(1´) [[I]$_{\text{NP}}$ strongly [advised]$_{\text{Verb}}$ [her]$_{\text{NP}}$ [[against]$_{\text{Prep}}$ [[[PRO]$_{\text{NP}}$ [undertaking this play]$_{\text{VP}}$]$_{\text{S2}}$]$_{\text{NP}}$]$_{\text{PP}}$]$_{\text{S1}}$

Structure (1´) makes use of the traditional idea of a nominal clause, that is, the lower clause, being a gerund, is dominated by an NP node. See Rosenbaum (1967) on the distinction between sentential complements that are nominal and those that are not (see also Rudanko 2011: 156).

Two general properties of object control constructions may be pointed out here. First, as regards the semantics of object control, the analysis of Sag and Pollard (1991) is the generally accepted view. Labelling matrix verbs selecting object control "verbs of the *order/permit* type", they write:

> Verbs of the *order/permit* type all submit to a semantic analysis involving STATES OF AFFAIRS (SOAs) where a certain participant (the referent of the object) is influenced by another participant (the referent of the subject) to perform an action (characterized in terms of the soa denoted by the VP complement). The influencing participant may be an agent (as in *Kim*

> *persuaded Sandy to leave*) or a nonagent (as in *Ignorance of thermodynamics compelled Pat to enroll in a poetry class*). The semantics of all verbs in this class thus involves a soa whose relation is of the INFLUENCE type. With respect to such soas, we may identify three semantic roles, which we will refer to as INFLUENCE (the possibly agentive influencer), INFLUENCED (the typically animate participant influenced by the influence) and SOA-ARG (the action that the influenced participant is influenced to perform (or, in the case of verbs like *prevent* and *forbid*, NOT to perform). [Note omitted] (Sag & Pollard 1991: 66)

Here the more traditional labels of Agent, Patient, and Goal are employed for the three semantic roles, but the labels Influence, Influenced, and SOA-ARG are richly descriptive and applicable to the analysis of *advise* in a sentence of the type of (1), and they also deserve to be kept in mind as more specific terms.

Another general property of object control in English has been labelled 'Bach's Generalization' in the literature, going back to Bach (1980: 304). To introduce the generalization, consider the sentences in (2a–d), from Rizzi (1986: 503):

(2) a. This leads people to the following conclusion.
b. This leads to the following conclusion.
c. This leads people [PRO to conclude what follows].
d. *This leads [PRO to conclude what follows].

Sentence (2c) involves object control, selected by the matrix verb *lead*, with an infinitival lower clause. Sentence (2d), which also has a lower clause with PRO, is ill-formed, whereas (2b), which does not have a sentential complement, is well formed. Bach's Generalization has been formulated to explain contrasts such as the one between (2d) and (2b). Rizzi (1986) provides a particularly clear definition of the generalization:

> In object control structures the object NP needs to be structurally represented. (Rizzi 1986: 503)

The contrast between (2b) and (2d) shows that the generalization applies to the matrix verb *lead*. However, as observed by Rizzi, the generalization is not a linguistic universal, and even in English there are certain exceptions to it. One class of exceptions concerns the matrix verb *warn* when used with sentential complements involving the preposition *against*. Consider the sentences in (3a–b):

(3) a. I would warn her against paying exorbitant prices for books and objects of art. (COHA, 1922, FIC)
b. Former U.S. Ambassador to Saudi Arabia Wyche Fowler warns against assuming that "monarchs can do anything they want without consequences [...]" (COHA, 2001, MAG).

In sentence (3a) the matrix verb *warn* selects the complementation pattern of NP [*against*]$_{\text{Prep}}$ [PRO Verb+*ing* ...]$_{\text{S}}$]$_{\text{NP}}$, with the higher object controlling the lower subject. The NP that immediately follows *warn* is an argument of the matrix verb, and the pattern is one of object control. In view of Bach's Generalization, the direct object would not be omissible, but the well-formedness of (3b) shows that *warn* permits exceptions to the Generalization. Rizzi (1986) is a pioneering study of such exceptions, with a focus on Italian and French. Taking account of Rizzi's work, exceptions to Bach's Generalization may be analysed as involving a covert or understood object NP in the higher clause that then controls the reference of PRO, the lower subject. The present study has the broad objective of contributing to work on Bach's Generalization.

2 Earlier work on exceptions to Bach's Generalization

Bach's Generalization was originally formulated within theoretical linguistics, without reference to, or use of, corpus data. Rizzi's (1986) study was likewise conducted within theoretical linguistics. However, the advent of large electronic corpora has opened up new avenues for investigating the Generalization and exceptions to it. Rudanko and Rickman (2014) and Rudanko (2015: Chapter 8) are earlier studies of this area of English grammar that make systematic use of corpus evidence. The former study investigated the matrix verb *warn* in each decade of COHA, with the focus on both overt and covert tokens of object control with this matrix verb in American English in the last two centuries, as in (3a–b). The study showed that the overt pattern has occurred in English throughout this period, with normalized frequencies ranging from 0.4 to 1.1 per million words in the different decades. The study also found that the covert object control pattern with *warn* was extremely rare in the nineteenth century, with no tokens at all in five decades of that century. However, the study also revealed that in the twentieth century the frequency of the covert pattern began to rise, especially from the 1930s onwards, and that in the 1990s and 2000s the covert pattern had in fact become more frequent than the overt pattern. When these two decades are combined, the numbers of tokens were 28 for the overt pattern and 38 for the covert pattern. Apart from such descriptive findings, the article also offered qualitative comments, derived from the tokens, on the interpretation of understood objects in violations of Bach's Generalization.

Rudanko (2015: Chapter 8) examined exceptions to Bach's Generalization in sentential complements of the matrix verb *counsel* in each decade of COHA and in two decades of the Corpus of Contemporary American English (COCA; M. Davies 2008–), the 1990s and the 2000s. With this verb both overt and covert object control constructions were found to be extremely rare in COHA, but the evidence of COCA showed that in the most recent decades the frequency of covert tokens (14) surpassed that of overt tokens (4). The numbers are low here, making qualitative comments hard to substantiate, but similarly to the more robust numbers of tokens in the case of *warn* in the most recent decades of COHA, these totals suggest that the covert pattern may be becoming more prominent in very recent English. Overall, these earlier studies suggest that there may be an emerging change in recent English favouring exceptions to Bach's Generalization, making it desirable to conduct more research in this area. This serves to motivate the present investigation.

The present study investigates the verb *advise* against the background of Rudanko and Rickman (2014) and Rudanko (2015). As in the earlier studies, the focus is on sentential complements introduced by the preposition *against*[1]. It is observed that alongside object control constructions with overt controllers, *advise* also permits covert controllers, in violation of Bach's Generalization. Even in the case of example (1) it is possible to omit the direct object on the main clause, to form the sentence in (4):

(4) I strongly advised against undertaking this play.

Comparing the sentences in (1) and (4) from the point of view of transitivity, the former, with an overt controller and an overt direct object of the main verb, is obviously more transitive than the latter, which lacks such an overt object. From this perspective, this study aims to contribute to work on a type of detransitivization. Most of the work on detransitivization (e.g., Garcia Velasco & Portero Muñoz 2002; Groefsema 1995) has concentrated on analyzing understood objects in simple sentences, but the present study, with its focus on Bach's Generalization, examines the omission of objects in front of sentential complements.

[1] The verb *advise* also selects a number of other patterns of sentential complementation, but these are left for later investigation.

3 Data and method

In order to investigate the incidence and the properties of object control constructions involving covert zero controllers of PRO in relation to overtly controlled complements with the verb *advise,* relevant tokens of the patterns were searched for in large electronic corpora. The time span of the study is that of the Corpus of Historical American English (COHA), that is from the 1810s to the 2000s. Some additional data from the British National Corpus (BYU-BNC version; M. Davies 2004–) and the Corpus of Contemporary American English (COCA) for present-day English are also considered. These corpora were selected because of their size, in order to examine a sufficiently representative sample of the two specific constructions at the syntax semantics interface. The corpora are also appropriate because of their carefully defined overall structure, with data from different text types, making it possible to study the constructions in question from the point of view of the system of language[2].

The task of choosing the appropriate search string tends to be more difficult in the study of object control than in the study of subject control, simply because of the more complex structure involved, but for the present investigation, the investigators adopted the string consisting of "[advise].[v*]," with the word *against* within nine words to the right. In order to guard against any errors in tagging in the corpora, a secondary search string consisting of "advis*" followed by *against* within nine words to the right was also used. These search strings are not ideal from the point of view of precision, but are designed to ensure maximum recall. In the COHA data, approximately 25 per cent of the tokens retrieved were relevant, and in the COCA data analysed, the corresponding percentage was around 30. Irrelevant tokens were excluded from further analysis.

The relevant tokens are examined first as regards the relations between the frequencies of the overt and covert object control patterns. The results gleaned from the diachronic COHA corpus are compared with those from the BNC and COCA. In addition to gerundial complements, attention is also given to the diachronic developments in the occurrence of overt and covert object control patterns when the matrix verb is followed by a noun phrase complement. The results of the quantitative analyses are presented in Section 4. The tokens are then analysed in a more qualitative fashion in order to investigate whether any semantic

[2] It might, in fact, be fascinating to consider the relevance of different text types on the question of the use of the two patterns. However, the numbers of occurrence in the data do not warrant making far-reaching conclusions on the issue.

and/or pragmatic characteristics can be observed that might play a role in the use of the two patterns. In this analysis, as presented in Section 5, earlier work on the interpretation of understood objects serves as the starting point.

4 The emergence of the covert object pattern

The results obtained with the search strings are given in Table 1, with the normalized frequencies given in parentheses:

Tab. 1: The incidence of overt and covert object control patterns with *advise* in COHA (PMW = frequency per million words).

DECADE	SIZE (million words)	OVERT TOKENS (PMW)	COVERT TOKENS (PMW)
1810s	1.2	0	0
1820s	6.9	0	0
1830s	13.8	3 (0.2)	0
1840s	16.0	3 (0.2)	1 (0.1)
1850s	16.5	0	0
1860s	17.1	0	0
1870s	18.6	2 (0.1)	0
1880s	20.9	2 (0.1)	1 (0.0)
1890s	21.2	0	0
1900s	22.5	1 (0.0)	4 (0.2)
1910s	22.7	2 (0.1)	2 (0.1)
1920s	25.6	7 (0.3)	1 (0.0)
1930s	24.6	2 (0.1)	2 (0.1)
1940s	24.1	7 (0.3)	2 (0.1)
1950s	24.4	6 (0.2)	5 (0.2)
1960s	24.0	7 (0.3)	0
1970s	23.8	2 (0.1)	2 (0.1)
1980s	25.2	4 (0.2)	9 (0.4)
1990s	27.9	4 (0.1)	9 (0.3)
2000s	29.5	4 (0.1)	5 (0.2)

Here are two initial illustrations of each pattern from COHA, one example from an earlier and another from a more recent part of the corpus:

(5) a. I was going to advise you against wedding a very poor girl. (COHA, 1835, FIC)
b. I advise students against taking teaching assistantships and research assistantships in their first year [...] (COHA, 1995, NEWS)

(6) a. He referred the matter to Neander, who advised against prohibiting it, with the assurance that [...] (COHA, 1843, FIC)
b. Public health departments advised against eating Great Lakes fish regularly, if at all. (COHA, 1998, NF)

The figures in Table 1 show that the covert object control pattern was extremely rare in the nineteenth century. The overt object control pattern was slightly less rare, but its frequency was also very low. From the 1920s onwards the overt pattern becomes slightly more frequent. As regards the covert pattern, there are some tokens found here and there in some decades, including five tokens in the 1950s, but for most decades its frequency is lower than that of the overt pattern. In the 1980s and 1990s the covert pattern rises in frequency, becoming noticeably more frequent than the overt pattern. To represent these trends more clearly, it is helpful to compare the incidence of the two patterns during thirty-year periods, beginning with the 1830s, in Table 2:

Tab. 2: The incidence of the overt and covert object control patterns with *advise* over thirty-year periods in COHA.

PERIOD	SIZE (million words)	OVERT TOKENS (PMW)	COVERT TOKENS (PMW)
1830–1859	46.3	6 (0.1)	1 (0.0)
1860–1889	56.6	4 (0.1)	1 (0.0)
1890–1919	66.4	3 (0.0)	6 (0.1)
1920–1949	74.2	16 (0.2)	5 (0.1)
1950–1979	72.1	15 (0.2)	7 (0.1)
1980–2009	82.5	12 (0.1)	23 (0.3)

As is evident in Table 2, the last two 30-year periods show a change in the ratios between the frequencies of overt and covert object control patterns, with the tokens of the covert object pattern outnumbering those of the overt objects in 1980–

2009. A simple Chi-Square test applied to the last two 30-year periods shows that the difference is statistically significant (p<0.05).

Given the evidence of COHA regarding the rising frequency of the covert object control pattern in very recent English, it is of interest here also to examine COCA, the Corpus of Contemporary American English, to see if the evidence of COHA is confirmed in this very large corpus of recent American English. A period of ten years from 2003 to 2012 was chosen for this comparison, in order to gain information on the latest trends in the use of the two constructions (see Table 3). The frequencies of the two patterns in Table 3 confirm that the covert object control pattern has indeed established itself in current American English, and it now clearly outnumbers the overt pattern.

Tab. 3: The overt and covert object control constructions with *advise against -ing* in COCA, 2003–2012.

PERIOD	SIZE (million words)	OVERT TOKENS (PMW)	COVERT TOKENS (PMW)
2003–2007	103.5	11 (0.1)	23 (0.2)
2008–2012	91.7	8 (0.1)	16 (0.2)
2003–2012 (total)	195.2	19 (0.1)	39 (0.2)

As regards the situation in British English, similar searches were conducted on the BNC, which covers the period between 1960 and 1993 (most of the data, however, dates from 1985–1993). In the entire BNC, containing 98.3 million words, 19 tokens were found of the overt object control construction with *advise against -ing* (0.2 per million words), and 16 tokens were found of corresponding constructions with a covert object (likewise rounding up to 0.2 instances per one million words). The earliest relevant search hits date from 1978, and the most recent ones from 1992. Here are some illustrations of both the overt and covert object control patterns with *advise against -ing* in the corpus:

(7) a. 'Big Jack' advised them against travelling to cheer on their team in a vital World Cup qualifying match, because of the primitive conditions in the Albanian capital, Tirana. (BNC, HJ4)
b. 'We don't know yet whether the young players of today will have the same problems,' adds Sortland, who, despite the results, will not be advising children against taking up football. (BNC, AKE)

c. Detectives say a number of forged tickets are in circulation and advise against buying tickets from touts. (BNC, K1H)
d. They claimed a report advising against fighting a costly appeal with the electricity giant should have been discussed in closed session. (BNC, K97)

Considering the frequencies of the overt and covert object tokens in American and British English, it would appear that the numbers of tokens of the two patterns are more even in the BNC data, while the figures from COHA representing roughly the same period are more suggestive of the objectless pattern outnumbering the cases with an explicit object. In light of this evidence, it could be suggested that the rise of the covert object control pattern with *advise* occurred earlier in American English. However, the study of larger sets of data would be needed to draw more definite conclusions.

In general terms, the rise of the covert object control pattern with *advise* may be one aspect of the increasing prominence of *-ing* complement clauses, which is an important aspect of what has been called the Great Complement Shift in some recent work on complementation (M. Davies 2012; Fanego 2016; Rohdenburg 2006, 2014; Rudanko 2012, 2017: 15–16, 29; Vosberg 2006, 2009). With respect to the Great Complement Shift, Rohdenburg (2006: 143) notes that "perhaps the most important set of changes is provided by the establishment of the gerund as a second type of non-finite complement", and because the covert *against -ing* pattern with *advise* adds to the arsenal of potential gerundial complements, it seems possible to view it as a manifestation of the Great Complement Shift[3].

With regard to the increase of the covert object control pattern with *advise*, another point worth considering is that there are also many instances where instead of a gerundial complement, *advise (NP) against* is followed by a non-sentential noun phrase, which may contain a noun or a pronoun as the head of the phrase, as in examples (8a–d) from COHA:

(8) a. 'Yes, we -- all her children -- think it's absurd. And we're all trying to advise her against it... but she vows she's going to get married to him anyhow.' (COHA, 1922, FIC)

[3] Rohdenburg (2006) also makes a more specific comment on *advise*, comparing the covert *to* infinitival pattern, as in *She advised to do it in advance*, with the non-prepositional gerund, as in *She advised doing it in advance*, and noting that the former has become rare, while the latter has become more frequent. The covert *against -ing* pattern of covert object control, which is the subject of the present study, cannot of course be directly compared to the covert *to* infinitival pattern because of the difference in meaning, but its emergence and increasing prominence is still in harmony with the Great Complement Shift.

b. 'Here, Sanders,' put in Prescott Gates. 'I'll handle this situation. I'll jolly well make them come out, whoever they are.'
'I strongly advise against it, sir,' said Sanders. (COHA, 1932, FIC)
c. 'I advise you against this trip, Mr. Taber,' said the manager. (COHA, 1922, FIC)
d. In any case he felt it would be both dangerous and useless to return to town unwell; and Feliu, observing his condition, himself advised against the journey. (COHA, 1889, FIC)

In examples (8a) and (8b), the pronoun *it* anaphorically refers to an explicit or implicit clausal argument, and in both cases the pronoun could be rephrased with *advise (NP) against* followed by an *-ing* complement (i.e., "we're all trying to advise her against getting married to him; I strongly advise against making them come out"). In a similar fashion, in the case of examples (8c) and (8d), one can consider that a verb in *-ing* has simply been omitted, corresponding with *making this trip/journey*. Furthermore, there are a number of tokens with either overt or covert objects where it is possible to regard the following noun as a nominalization of a verb of action or process, as in examples (9a–b) from COHA, in which the nouns *use* and *enlargement* could also be expressed with the *-ing* forms *using* and *enlarging*:

(9) a. He advised citizens against the use of drugs to combat high blood pressure because, he said, it led to impatience. (COHA, 1983, MAG)
b. Justice Lazansky declined to comment on a report that he had advised against the enlargement of the investigation. (COHA, 1931, NEWS)

Tab. 4: The incidence of the overt and covert objects with *advise against* followed by a noun or noun phrase over thirty-year periods in COHA.

PERIOD	SIZE (million words)	OVERT TOKENS (PMW)	COVERT TOKENS (PMW)
1830–1859	46.3	1 (0.0)	0 (0.0)
1860–1889	56.6	4 (0.1)	15 (0.3)
1890–1919	66.4	11 (0.2)	21 (0.3)
1920–1949	74.2	15 (0.2)	37 (0.5)
1950–1979	72.1	6 (0.1)	25 (0.3)
1980–2009	82.5	11 (0.1)	30 (0.4)

It is, then, interesting to observe the occurrences of overt and covert tokens with *advise (NP) against* + a noun phrase in the corpus data. Table 4 presents the frequencies of the two patterns in COHA. The general trend as regards the higher frequencies of instances with covert object tokens is similar to that seen in Table 2. However, the increase in the occurrences of covert object tokens in this case established itself rather early. While some of the earliest COHA instances have overt tokens, objectless instances soon outnumbered them.

With regard to the overt and covert object patterns with *against* + NP complements in COCA, based on the 2003–2012 data, covert objects have become even more frequent compared to the overt ones, with altogether 10 instances found with an explicit object, and as many as 66 instances without an object. In the BNC, the corresponding figures are in line with the corpora representing American English, with 16 tokens with overt objects, and 43 tokens with covert objects.

Based on corpus evidence presented in Tables 2 and 4, it would seem justifiable to argue that the well-established use of objectless *advise* when followed by an *against* + NP complement helped pave the way for the later increase of the objectless *advise against -ing* constructions. In fact, a comparison between the figures in Tables 2 and 4 shows that the normalized frequencies of the covert object control pattern with sentential complements (0.3 per million words) are almost beginning to rival those with corresponding NP complements (0.4 pmw). A point that remains open concerning the possible connection between the two patterns is why the increase of covert objects with the gerundial *against -ing* did not begin until the 1980s if covert objects were already used with NP complements in the late 19th century. In any case, the emergence of covert tokens appears evident, which raises the question of whether there are any special nuances perceivable in the usage of the more recent pattern which would justify its rise alongside the earlier overt object control pattern. This issue is discussed further in the light of corpus examples in the following section.

5 Semantic and pragmatic considerations in the use of the covert object pattern

The numbers of tokens of the covert object control pattern in the corpora examined are numerous enough to permit some discussion of the properties of the pattern. The focus here is on data from COCA, in order to shed light on usage that is as current as possible. The discussion is presented here in three separate parts, according to the different semantic and pragmatic considerations which can be seen as contributing to the use of the pattern.

5.1 Interpretations of habituality and actualization

A suitable point of departure for the closer study of the covert patterns is provided by some comments in Huddleston and Pullum (2002: 303) on unexpressed objects with selected classes of verbs. One of the types in that source is labelled "Unexpressed human object", as in *The dog bites*. In discussing a class of salient verbs, the authors observe that some verbs selecting unexpressed human objects, including *please*, "appear more readily in intransitives when the situation is habitual or unactualized – e.g. *He never fails to please, I'll aim to please,* but hardly *?His behavior at lunch pleased*" (Huddleston & Pullum 2002: 303). These comments suggest two dichotomies: one between habitual versus one-off situations and another dichotomy between actualized and unactualized ('irrealis') situations. These dichotomies are not presented as relevant to understood objects in the case of *advise* in Huddleston and Pullum (2002: 303), but it is still worth considering them in the context of the covert object control pattern. The primary focus of the study being on present-day English, the 39 tokens in the most recent ten year period (2003–2012) of COCA may again serve as the main data for analysis, supplemented by some earlier tokens from the corpus as clarifying examples. The covert object control pattern is more complex than the omission of a direct object with a simple transitive verb, as in *The dog bites*, and it is possible to consider the application of the dichotomies to the sentential complements of *advise*.

The salient sense of *advise* is 'to give guidance or suggestions' (OED, sense 5), and it is in principle possible to give guidance against a course of action that is open to the addressee or that the addressee is contemplating, but has not yet engaged in at the time when the advice is given, but it is also possible to give guidance against a course of action that the addressee has already engaged in. Examples of the latter kind include the following instances from COCA (10a–b):

(10) a. Senator Hagel: Burning tapes, destroying evidence, I don't know how deep this goes. […]
Bob Schieffer: Reports say that Harriet Miers, who was then a lawyer on the president's staff in the White House, found out about this and told the CIA, advised against doing it. And apparently they went against those orders. (COCA, 2007, SPOK)
b. At least 79 of the suspects stunned in 2009–2010 had offered either passive, verbal resistance or were fleeing an officer, reports show. However,

guidelines issued by a national police research group and the U.S. Justice Department office of community policing in 2005 advised against stunning suspects who are simply fleeing police. (COCA, 2011, NEWS)

In (10a), it becomes evident from the context that the action advised against has already been taken, possibly on several occasions. In (10b), it is not overtly stated that the action referred to in the lower clause had taken place before the advice was given, but this interpretation is possible. The question of habituality may also arise: in examples (6b) and (10b) the advice may be read as advice against a habit (of eating Great Lakes fish or of stunning suspects), but it is also possible to read the sentences as offering advice against even a single (future) act (of eating Great Lakes fish or of stunning suspects).

Although the context in some instances found in the data suggests that the action advised against had already taken place with indeterminate degrees of regularity or habituality, for the majority of the tokens, irrealis interpretations seem relevant, and in such cases the question of a habit obviously does not arise. For instance, in *Mr. President, I strongly advise against sending federal troops to Chicago* (COCA, 1996, FIC), the construction suggests an irrealis interpretation in that it does not convey that federal troops had already been sent to Chicago. Here are some further examples of irrealis readings in sentences not illustrated above:

(11) a. They had demonstrated no contrition. Two U.S. attorneys who prosecuted them advised against granting clemency. (COCA, 2009, SPOK)
b. To get a temporary commitment order from a judge, the state must present two medical recommendations. One psychiatrist who supplied them, Dr. Gerald Groves, said that sometimes, if he advised against committing someone, 'the institution might go find another psychiatrist who would be willing to commit.' (COCA, 2003, NEWS)

It is worth noting that the analysis of the tokens along the lines of "actualized vs. irrealis" is not always a straightforward matter, but that there are occasionally further complexities involved. Good cases in point can be observed in examples (12a–b):

(12) a. The committee found there was no specific prohibition in the Koran on driving. In fact, during the time of the Prophet, women regularly led camels across the desert. Even now, Bedouin women have regularly been permitted to drive cars and trucks in isolated parts of the kingdom. The committee nevertheless gently advised against repeating the experiment. (COCA, 1990, MAG)

b. The celebrities worked to bridge the gap between their world and that of their audience – some with more success than others. Monica, the Grammy-award winning R&B singer, told about her struggles with credit cards and advised against signing up for more than one card or charging big-ticket items. (COCA, 2007, NEWS)

Example (12a) is interesting in that the action advised against had taken place before the committee gave its advice, and this action was even done on a regular basis. However, what complicates the interpretation of this token is that the verb in the lower clause (what is being advised against) is *repeat*, which refers to future actions as well as past ones. Example (12b), on the other hand, presents another type of borderline instance in that the actualized situation (of signing up for more than one credit card) involves the subject rather than the understood NP object of *advise*, and while the context suggests that the addressees may have been struggling financially, it does not imply that they actually had signed up for more than one credit card. Overall, the present investigators counted 32 tokens with irrealis interpretations of the sentential complement of *advise* among the 39 tokens in the COCA material. The actions advised against in the remaining seven tokens, including examples such as (12 a–b), were interpreted as having already occurred[4].

5.2 Degree of specificity of the object

Turning to a second perspective on the properties of the covert control pattern, we may take note of another comment by Huddleston and Pullum (2002). Writing on a class of verbs with implicit objects that includes *advise*, they observe:

> We interpret the intransitives as having a human object, but it may be either general (arbitrary people), as in *That dog bites*, or specific, e.g. *you* in particular, as in a salient interpretation of *Take care; it may bite*. (Huddleston & Pullum 2002: 303)

The contrast between general versus specific interpretations of the implicit object is of interest and clearly deserves comment. The contrast links to Rizzi's (1986)

[4] A construction with an object that is implicit is clearly lower in transitivity than a construction with an expressed object (Hopper & Thompson 1980), and it may be suggested that the incidence of irrealis interpretations goes well with the reduced transitivity or detransitivization that is implicit in the covert object construction. However, further examination with a larger dataset would be required to establish whether there indeed is such a connection.

discussion of covert controllers in object control constructions, for in that discussion the role of a general interpretation of the understood object is prominent. Rizzi, with a focus on Italian and French, more than on English, did not make use of corpus evidence, and we therefore examine some authentic tokens here from this perspective. COCA is again the main source of data here, but some early examples from COHA are worth considering first:

(13) a. When Strauss's work was published, it was proposed to his majesty, the King of Prussia, to prohibit it in his dominion. He referred the matter to Neander, who advised against prohibiting it, with the assurance that it would make eventually more apparent the divine origin of the religion it so powerfully attacks. (COHA, 1843, FIC)
b. He is not parsimonious, but his instincts and habits have been prudent. He is making inroads upon his capital, and if he should never get it back? His father, it is true, advised against entangling his private fortune, but it can not be helped now. (COHA, 1883, FIC)
c. Some high powered gasoline cars are well calculated for touring, and are comfortable with four or five in the party, but I strongly advise against overloading any car […] (COHA, 1905, MAG)

As noted, the numbers of early tokens are few, but even these illustrations serve to show that both specific and general interpretations must have been available in the interpretation of the understood NP from early on. Thus in (13a–b) the interpretation seems specific, confined to one individual, with the reference supplied by the context (*the King, his*) but in (13c), the interpretation is general, referring more broadly to any people.

To shed more light on the nature and interpretation of the understood NP, it is helpful to turn to the more recent times, where examples are more plentiful. The 39 tokens of the covert construction found in the COCA data (2003–2012) are examined here more systematically. Here are some illustrations from this period:

(14) a. Since you already own Labrador retrievers, I'd advise against adopting a potbellied pig. Both are pack animals by nature […] (COCA, 2009, MAG)
b. There are no exact parallels, of course, but I suspect your attorneys will advise against revealing what you have uncovered, just as they'd advise against revealing anything the government discovers in its normal operations. (COCA, 2012, FIC)

(15) a. While baggage ID tags are essential for reuniting lost luggage with their owners, security experts advise against putting your home address on these. (COCA, 2005, NEWS)

b. [...] they're [decongestants] no longer recommended for kids under 2, and many doctors advise against giving them to older kids. (COCA, 2008, MAG)
c. Experts advise against letting a ruling party or leader tinker with the state's basic architecture. (COCA, 2011, NEWS)

The examples in (14a–b) testify to the possibility of highly specific interpretations of understood objects. For instance, in (14a) the understood object may be taken to refer to the referent of *you*. By contrast, in (15a–c) the NP is understood in a broader way, and extends beyond one individual. For example, in (15a) the possessive *your* probably refers to the public in general, and in (15b) the plurals in the sentence are worth noting, likewise suggesting a fairly broad or general interpretation of the understood object of *advise*.

There are also several tokens where the reference of the understood NP is less specific or less definite than in (14a–b), but also less general than in (15a–b):

(16) a. Colby had briefly been assigned to a unit in Nha Trang before coming to Phu Hiep, and I wrote an official sounding report advising against charging Colby with selling a box of hand grenades to suspected enemy agents because there was insufficient evidence. (COCA, 2009, MAG)
b. After talking things over, they'd give their best advice to Congress and to the president. It wouldn't be compulsory for them to take it, but the board's opinion would be made public and that would put pressure on politicians. For example, if the professors had advised against attacking Iraq, it would have been harder for the president to do that. The professors would give Congress and the president their best advice on whether or not they should pass a new tax decrease that leaves no rich person behind. (COCA, 2004, SPOK)

In (16a–b) the reference of the understood NP is limited by the lower predicates of these sentences to well-defined subsets of individuals, military prosecutors in (16a) and Congress and the president in (16b). Similarly, the understood NP in example (10b) cited earlier may be regarded as referring to police officers.

5.3 Pragmatic considerations

There is a further point to be made regarding the interpretation of the understood NP. Consider the illustrations in (17a–b):

(17) a. The next moment, they were in the hallway, hugging each other as if nothing else mattered. ‘My dearest Mrs. Talbot.’ Lenore pulled away before she broke into tears. ‘My sister – ??’ ‘Sleeping at last, poor lamb. The apothecary was here this afternoon and says she is very bad, very low indeed. He advised against bleeding her again – but it is not my place to say even that much. I plead an old woman’s errant tongue.’ (COCA, 2011, FIC)
b. Some of Obama’s aides counseled against giving a speech that focused directly on race. ‘I know there were some who advised against making this speech,’ says Rep. Artur David, an Alabama Democrat and Obama supporter. (COCA, 2008, NEWS)

In (17a) the advice is presumably given to a small and specific subset of individuals – Mrs. Talbot and her friends or relatives – and a similar interpretation seems appropriate in (17b), though in the latter case the set of recipients may have been larger. What makes (17a) and (17b) interesting is that they suggest that in the covert pattern there may be some indeterminacy regarding the interpretation of the relation of the understood object of the verb *advise* and the understood subject of the lower clause, represented by PRO. In the discussion up to this point it has been assumed that PRO is controlled by the understood object and that the pattern is one of straightforward object control. It is recalled that the very definition of Bach’s Generalization makes reference to object control. However, (17a) and (17b) suggest that in the covert pattern there may be some indeterminacy in the interpretation of PRO. The possibility of indeterminacy is clearer in (17b). The passage is about a speech delivered by Obama, and the understood subject of *making this speech* refers to Obama. It is also possible to interpret the understood object of *advise* as referring to Obama. However, the sentence can also describe a scenario where the advice was not given directly to Obama but to some of his aides or to his chief of staff for instance, at least by some of the advisors who had offered the advice. This interpretation of (17b) may be called the mediated or proxy reading of (17b). Even in the proxy reading the referent of the understood object and the referent of the understood lower subject are understood to be connected, but the NPs in question are not coreferential.

The possibility of a proxy reading in (17b) is of interest because if an overt object NP is inserted into (17b), there is less possibility of indeterminacy: a sentence such as *I know there were some who advised the chief of staff against making this speech* suggests a straightforward object control interpretation more strongly as a default reading. The possibility of indeterminacy in the case of the covert pattern functions as a niche for semantic differentiation between the two patterns. A similar indeterminacy may arise in (17a), for the sentence appears to permit a scenario where the apothecary’s advice was given to the sick woman’s friends or relatives, while the actual bleeding, if it had been carried out, might have been carried out by a doctor, as instructed by the friends or relatives.

The irrealis versus actualized dichotomy, discussed above, raises a question for the interpretation of the implicit object. As noted, the overwhelming majority of tokens is of the irrealis type, and it is observed that irrealis interpretations are often found with specific implicit objects. Thus both of the specific implicit objects in (14a-b) involve irrealis readings of the lower clause. Further irrealis readings can also be found with implicit objects that are less specific and more indeterminate, as for instance in (16a).

Regarding the reason(s) for the use of the covert object control construction, in Rudanko and Rickman (2014) two suggestions were made. As a number of tokens of the construction with *warn against -ing* were found in texts dealing with politics, it was postulated that the very possibility of indeterminacy of the interpretation of PRO might make the construction attractive in political discourse. However, in the present study the tokens of the corresponding construction with *advise* in COHA and COCA were found in texts concerning a variety of topics, and politics as a subject matter was not as prominent among the tokens as it was with *warn*. It is nevertheless possible that verbs within the same semantic field even in this case would show some general similarities in their selection of complements, while reasons for why occurrences of *advise against -ing* are not particularly prominent in political discourse in the same way as *warn against -ing* can perhaps be found in the nature of the discourse itself. For example, one could argue that in political discourse those lines of argumentation are generally preferred which are characterized by certainty, strength, and assertiveness rather than tentativeness, restraint or caution, and *warn* being more assertive than *advise* could perhaps explain the prominence of *warn against -ing* in political discourse. The indeterminacy of PRO may play a role on a more or less conscious level in the use of the covert object control pattern in the case of both *warn* and *advise*, and possibly other semantically related verbs as well.

Rudanko and Rickman (2014) also made the suggestion that the omission of an explicit direct object may be linked to mitigating the potentially face threatening nature of a warning as a speech act, in a manner that may be similar to omitting the *by* phrase in passives (cf. Wanner 2009: 184). Warning someone against doing something may be more face threatening as a speech act than advising someone against doing something, but there may be some degree of imposition or face threat involved even in advising someone, or in taking it upon oneself to advise someone, because even advising involves influencing the behaviour of another person. The omission of the direct object may then sometimes be viewed from this perspective. For instance, consider (18) from COCA:

(18) Economics should concern itself with political and social needs, he argued, and he called for an end to the prize in economics. The free-market conservative Friedrich von Hayek, who shared the Nobel in economics with Myrdal in 1974 despite being his ideological opposite, agreed on that point. If he had been asked, Hayek said, he would have 'decidedly advised against' creating an economics prize. (COCA, 2007, NEWS)

The omission of the direct object in (18) is in a quotation, and in leaving the object indeterminate the speaker avoids addressing the question of who or what body of people should have acted differently at an earlier point in time, and in this way the potential face threat of the utterance of the speaker as a criticism is mitigated.

6 Conclusion

Earlier literature on complementation and other relevant grammatical structures offer many observations describing and explaining the changes in the uses of the structures. The characterizations can be examined more closely in connection with individual complement-taking predicates such as *advise* with the aid of large electronic corpora. It has been observed in this paper that based on the study of COHA and COCA, the construction *advise against -ing* is used with both overt and covert objects, thus violating Bach's Generalization. In fact, the latter type has become more frequent in recent decades, a development which becomes clear even in the fairly modest number of occurrences in COHA. Supporting evidence for the trend was found in COCA.

The possible role of two semantic features on the choice of leaving out the object in the *advise against -ing* structure was considered, namely the "actual or irrealis" nature of the situation or activity expressed in the lower clause, and the degree of specificity of the covert object. The data showed that while different kinds of occurrences can be found in the data regarding the two dichotomies, covert objects are more frequent when the situation is of the irrealis type and the understood object is either general or indeterminate. As was seen in connection with the structure *warn against -ing*, in certain circumstances it may be beneficial to the speaker to use a construction where the interpretation of the covert object is to some extent indeterminate. It could be postulated that *advise against -ing*, as well as other semantically related verbs, may show similar tendencies in this regard, which is a question that deserves further study.

References

Bach, E. (1980). In defense of passive. *Linguistics and Philosophy*, *3*(3), 297–341.

Carnie, A. (2007). *Syntax. A Generative Introduction*. (2nd ed.). Malden, MA: Blackwell.

Chomsky, N. (1984). *Lectures on Government and Binding. The Pisa Lectures*. (3rd rev. ed.). Dordrecht, Netherlands: Foris.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York, NY: Praeger.

Davies, W. D., & Dubinsky, S. (2004). *The Grammar of Raising and Control*. Malden, MA: Blackwell.

Davies, M. (2004–). *BYU-BNC* (Based on the British National Corpus from Oxford University Press). Retrieved from http://corpus.byu.edu/bnc (accessed April 2014, last accessed October 2018).

Davies, M. (2008–). *The Corpus of Contemporary American English* (COCA): 520 million words, 1990-present. Retrieved from http://corpus.byu.edu/coca (accessed April 2014, last accessed October 2018).

Davies, M. (2010–). *The Corpus of Historical American English* (COHA): 400 million words, 1810-2009. Retrieved from http://corpus.byu.edu/coha (accessed April 2014, last accessed October 2018).

Davies, M. (2012). Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In T. Nevalainen & E. Traugott (Eds.), *The Oxford Handbook of the History of English* (pp. 157–174). Oxford, UK: Oxford University Press.

Fanego, T. (2016). The Great Complement Shift revisited. *Functions of Language*, *23*(1), 84–119.

García Velasco, D., & Portero Muñoz, C. (2002). Understood objects in functional grammar. *Web Papers in Functional Grammar,* 76. Retrieved from http://home.hum.uva.nl/fg/working_papers/WPFG76.pdf (last accessed October 2018).

Groefsema, M. (1995). Understood arguments: A semantic/pragmatic approach. *Lingua*, *96* (2–3), 139–161.

Hopper, P., & Thompson, S. (1980). Transitivity in grammar and discourse. *Language*, *56*(2), 251–289.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Jespersen, O. ([1940] 1961). *A Modern English Grammar on Historical Principles. Part V: Syntax,* volume IV. Reprinted 1961. London, UK and Copenhagen, Denmark: George Allen and Unwin/Ejnar Munksgaard.

Rizzi, L. (1986). Null objects in Italian and the theory of pro. *Linguistic Inquiry, 17*(3), 501–557.

Rohdenburg, G. (2006). The role of functional constraints in the evolution of the English complementation system. In C. Dalton-Puffer, D. Kastovsky, N. Ritt, & H. Schendl (Eds.), *Syntax, Style and Grammatical Norms* (pp. 143–166). Bern, Switzerland: Peter Lang.

Rohdenburg, G. (2014). On the changing status of *that*-clauses. In M. Hundt (Ed.*), Late Modern English Syntax* (pp. 155–181). Cambridge, UK: Cambridge University Press.

Rosenbaum, P. (1967). *The Grammar of English Predicate Complement Constructions.* Cambridge, MA: MIT Press.

Rudanko, J. (2011). *Changes in Complementation in British and American English*. Basingstoke, UK: Palgrave Macmillan.

Rudanko, J. (2012). Exploring aspects of the Great Complement Shift, with evidence from the *TIME* Corpus and COCA. In T. Nevalainen & E. Traugott (Eds.), *The Oxford Handbook of the History of English* (pp. 222–232). Oxford, UK: Oxford University Press.

Rudanko, J. (2015). *Linking Form and Meaning*: *Studies on Selected Control Patterns in Recent English*. Basingstoke, UK: Palgrave Macmillan.

Rudanko, J. (2017*). Infinitives and Gerunds in Recent English: Studies on Non-Finite Complements with Data from Large Corpora*. London, UK: Palgrave Macmillan.

Rudanko, J., & Rickman, P. (2014). Null objects and sentential complements, with evidence from the Corpus of Historical American English. In K. Davidse, C. Gentens, L. Ghesquière, & L. Vandelanotte (Eds.), *Corpus Interrogation and Grammatical Patterns* (pp. 209–221). Amsterdam, Netherlands: John Benjamins.

Sag, I., & Pollard, C. (1991). An integrated theory of complement control. *Language, 67*(1), 63–113.

The Oxford English Dictionary (OED Online). Oxford: Oxford University Press. Retrieved from http://www.oed.com (accessed April 2014, last accessed October 2018).

Vosberg, U. (2006). *Die Grosse Komplementverschiebung*. Tübingen, Germany: Gunter Narr.

Vosberg, U. (2009). Non-finite complements. In G. Rohdenburg & J. Schlüter (Eds.) *One Language, Two Grammars? Differences between British and American English* (pp. 212–227). Cambridge, UK: Cambridge University Press.

Wanner, A. (2009). *Deconstructing the English Passive*. Berlin, Germany: de Gruyter.

Magnus Levin

Subjective progressives in the history of American English

He's always telling some kind of lie

Abstract: This study investigates subjective progressives with *always/constantly/forever* (e.g., *you're always complaining*) in American English with data from the Corpus of Historical American English (COHA). The results show that subjective progressives are increasing but that this shift is restricted to *always* + progressive. The increase in subjective progressives is linked to colloquialization as reflected in an increasing use of first-person subject pronouns and contracted verb forms. Fiction contains the highest frequency of subjective progressives, largely due to fictional dialogue expressing subjective attitudes. The proportion of negative subjective attitudes decreases slightly towards the end of the 20th century. The material indicates that women are leading the way in the increase in subjective progressives, but there is no difference in women's and men's preferences for expressing negative subjective attitudes.

1 Introduction

Although the progressive in English has been studied extensively, there are still gaps in the research regarding low-frequency uses of the construction. Previous studies (see e.g. Killie 2004; Kranich 2010a; Leech et al. 2009: 118–143; Smith & Leech 2013; Smitterberg 2005) typically investigate the use and changing frequencies of the progressive in small, carefully sampled corpora while devoting only a few pages to different "special uses" (as termed by Leech et al. 2009: 119). One such special use concerns progressives occurring together with ALWAYS-type adverbials as exemplified in (1) and (2). These have so far not been studied in modern large-scale corpora, in spite of previous studies indicating that the absolute frequency of these constructions is on the increase. The rise of such progressives is connected to trends affecting English on a general level, such as colloquialization (Leech et al. 2009) and subjectification (Kranich 2007). To establish such connections, this chapter investigates the changing and variable patterns of

Magnus Levin, Linnaeus University, magnus.levin@lnu.se

https://doi.org/10.1515/9783110489071-011

progressives occurring with the adverbs *always*, *constantly* and *forever* in the Corpus of Historical American English (COHA).

(1) "Now wait a second, Hilda, until I've finished," he said testily. "You young people *are always interrupting*. [...]" (COHA; Fiction; 1953)

(2) He*'s always trying* to learn. He'll get up in the world. (COHA; Fiction; 1854)

Examples are given in (1) and (2) containing the most frequent adverbial, *always*. It is usually pointed out (e.g., Kranich 2010a) that speakers express their disapproval of something by using hyperbolic ALWAYS-type adverbials with progressives, as in (1), but examples such as (2), expressing intensification rather than disapproval, also occur. In (1) the negative attitude is indicated by the choice of the verb *interrupting*. This is strengthened by the adverbial *testily* in the narrative text. In contrast, (2) refers to a positive attitude attributed to the subject, a willingness to learn. What unites (1) and (2), however, is a hyperbolic, subjective component.

There is variation in the terminology with these progressives, which have been variously referred to as "modal" (Wright 1994), "subjective" (Killie 2004), "not-solely-aspectual" (Smitterberg 2005: 210–217) or "expressive or attitudinal" (Leech et al. 2009: 134). The present study opts for the term 'subjective' progressives because it highlights the connection to subjectification, i.e. the tendency for linguistic meanings to become more and more subjective (Traugott & Dasher 2001). Subjectivity in this sense refers to speakers' expression of self and the representation of perspective or point of view in discourse, and subjectification refers to the processes of linguistic evolution that lead to such strategies (Finegan 1995).

The spread of the progressive has also been studied from a sociolinguistic perspective. Investigations (Arnaud 1998; Smitterberg 2005) have found that women are leading the way in its overall increase. This study therefore also takes the language-external factor of gender into account when charting subjective progressives.

More specifically, the aims of this chapter are to explore

1. the diachronic development of *always/constantly/forever* + progressive in COHA,
2. the language-internal factors (such as colloquialization) affecting the diachronic and synchronic variation,
3. the extent to which there is ongoing subjectification in the sense of increasing negative speaker attitudes (annoyance/dislike/sarcasm) expressed by these progressives,

4. the extent to which there are gender differences in the use of the progressive in COHA (Arnaud 1998; Smitterberg 2005).

Section 2 discusses previous work on subjective progressives, while Section 3 presents the method and material used. Section 4 analyses the results, and Section 5 contains a concluding discussion of the findings.

2 On the history of the progressive verb in English

As mentioned above, a number of studies have noted an increase of the progressive in general, providing converging evidence for an increase over longer time spans. To be exact, Hundt (2004) found a fourfold increase in the absolute frequency of the progressive in American English (AmE) between the late 18th century and the late 20th century in the ARCHER corpus (A Representative Corpus of Historical English Registers; manchester.ac.uk/archer/), and Kranich (2010a) found similar results for British English (BrE) in the ARCHER-2 corpus. In Smitterberg's (2005) BrE material from the Corpus of Nineteenth-Century English (CONCE) (see Kytö, Rudanko, & Smitterberg 2000), there was also a clear growth of progressives of almost 50% per 100,000 words. This shift continues into the 20th century for both AmE and BrE (see, e.g., Kranich 2010a; Leech et al. 2009: 118–143; Smith & Leech 2013). Leech et al.'s (2009: 289) written AmE data from the Brown and Frown corpora indicate a significant 15% increase in progressives in fiction between 1961 and 1991, while there was a significant 21% decrease in the learned genre. The learned genre is generally slow to adopt changes, and this is also a genre where subjective attitudes are less expected due to its typically objective, impersonal style. It is therefore not surprising that learned writing and fiction are diverging in their preferences, but it is not obvious why progressives are decreasing in the learned genre.

The increase in the progressive is not restricted to writing, as shown by Leech et al.'s (2009: 126) finding of a 45% increase in the progressive in spoken BrE in the Diachronic Corpus of Present-Day Spoken English (DCPSE) between the 1960s and the early 1990s. Progressives were also found to be consistently more frequent in speech than in writing, which shows that genre and register are important factors in the distribution. In his 19th-century BrE data, Smitterberg (2005: 63) found that speech-related genres (such as trials, drama and letters) generally yielded many progressives, while expository genres (history, science) contained fewer. Fiction with its frequent use of dialogue covered the middle ground. These distributions have changed over time, as can be seen in a comparison with Leech et al.'s (2009) numbers. In the Brown family of corpora, the fiction sub-corpus contains the highest frequency of progressives both in AmE and BrE, well ahead of the newspaper texts. The learned texts produced only about one-fifth of the progressives found in fiction (for similar findings, see e.g. Biber et al. 1999: 460–463). In contrast, Smitterberg (2005) found relatively high frequencies of progressives in scientific texts from the 19th century. It thus appears that science was a slightly different, more involved genre in earlier centuries (see Taavitsainen 1997 for a discussion of the expression of personal affect in Early Modern English), and Smitterberg hypothesizes that the progressive possibly had less emotional force in the 19th century than in present-day English. This view is supported by Núñez-Pertejo (2004), who found low frequencies of subjective progressives in Early Modern English, therefore arguing that speakers only rarely used the progressive as an attitude marker in that period.

Earlier studies use different definitions and classifications of the phenomenon at hand, which makes it difficult to compare findings. Killie (2004) argues that the only certain way of identifying subjective progressives is to include merely those that occur together with ALWAYS-type adverbials. She further argues that ALWAYS-type progressives typically do not have aspectual features such as the activity described being in progress or temporary, and that the subjective meaning derives from a combination of the adverb and the progressive form (Killie 2004). Statements that someone is always doing the same thing are by definition hyperbolic and hence expressions of subjective perceptions (see further e.g. Visser 1973: 1942). Killie (2004: 44) suggests that the reason progressives are used in such contexts is that the progressive itself is at least partly subjective. Similarly, Smitterberg (2005: 210–217) includes such instances in his category 'not-solely-aspectual progressives', while Römer (2005) additionally includes instances without ALWAYS-type adverbials in her categories 'emphasis' or 'attitude'. Finally, Kranich (2010a: 214), based on Wright (1994), separates 'subjective' from 'objective' occurrences of progressives with ALWAYS-type adverbials, the defining criterion being "whether reference is made to a process which is not objectively ongoing at all times but where the speaker uses the adverb hyperbolically". Since the present chapter is restricted to and includes all instances of *always/constantly/forever*, the results are more directly comparable to studies only covering progressives co-occurring with adverbials.

Subjective attitudes with the progressive have previously been defined in different ways. Kranich (2010a: 63) points out that many scholars argue that there is an "inference of negative speaker attitude" with the progressive and adverbials such as *always*. Citing an adapted example from Leech (1987: 34), *Paul is always giving people lifts*, Kranich (2010a: 65) argues that such progressives tend to be interpreted as expressions of negative attitudes in spite of the habit of giving lifts usually being thought of as commendable. Kranich argues that the negative attitude here resides in the combination of the adverbial and the progressive, rather than in the adverbial alone. The present study broadly adheres to Kranich's (2010a) approach, but includes all instances of *always/constantly/forever* + progressive in the data (the 'neutral' instances being excluded from Kranich's category). This scope enables diachronic comparisons of the distributions of negative and positive/neutral connotations.

The diachronic developments of subjective progressives have been investigated extensively, but most of these studies are based on small-scale corpora with only relatively few instances. Killie's (2004) Early Modern English material indicated a preference for such progressives to occur in the present tense and with first- and second-person pronoun subjects. The correlation between subjective progressives and first- and second-person pronoun subjects has also been noted in synchronic studies. Römer (2005: 99–100) mentions that emphatic progressives tend to occur with first-person subjects and with the adverbials *always*, *now* and *all the time* in her synchronic study. When annoyance is expressed, subjects are fairly often second- and third-person pronouns. Römer thus conflates the categories emphasis and (negative) attitude in her study. In spite of the progressive increasing overall in his diachronic CONCE material, Smitterberg (2005) found no change in the percentage of subjective progressives in comparison with progressives overall across the 19th century. He argues that the reason for this is that subjective progressives have already been integrated into English grammar since Old English times. Kranich (2010a: 217) suggests that the tendency to use ALWAYS-progressives with negative subjective attitudes is a fairly recent, 20th-century development. Her data (2010a: 213–222) even indicate that subjective progressives – both those with ALWAYS-type adverbials and subjective progressives without such adverbials (Kranich 2008: 248) (and interpretative progressives) – may have been decreasing since Early Modern English times. Leech et al. (2009) found a slight increase in ALWAYS-type progressives between the 1960s and 1990s in the Brown family of corpora, but no far-reaching conclusions could be drawn from their low frequencies.

Finally, Laitinen and Levin (2016) investigated subjective progressives with *always/constantly/forever* in several different corpora: written AmE from the *Time* corpus (http://corpus.byu.edu/time/), consisting of 100 million words from all issues of *Time* Magazine 1923–2006, AmE speech from the five-million-word Longman Spoken American Corpus, lingua franca speech from the one-million-word Vienna-Oxford International Corpus of English (VOICE), and written English from Finland and Sweden from blogs, newspapers and student theses (one million words). The data indicate a slight increase in written AmE from *Time* towards the latter part of the 20th century. This change is at least partly connected to the colloquialization of writing habits. The findings also indicate that subjective progressives are more than twice as frequent in AmE speech as in magazines. The frequencies in the lingua franca material were in between the written and spoken AmE material. The diachronic material did not indicate any tendency towards subjectification in the sense of increased uses of negative attitudes, but rather a specialisation in that *always* + progressive is increasing while *constantly* and *forever* may either be on the decrease or remain stable. Semantic constraints seem also to be in operation, as *constantly* + progressive is becoming restricted to more neutral meanings while *always* and *forever* are becoming more closely associated with negative contexts.

In previous research, different extralinguistic factors have been connected to the increasing use of the progressive. One of these relates to the leading role of women in language change. Labov (2001: 292–3) summarizes the general findings regarding gender and language change thus: women typically use more of the innovative forms in changes from below, but they also adhere more strictly to overtly prescribed norms than men do. This latter tendency is probably linked to a desire to speak “properly” in order to achieve higher status in society, as proposed by Smitterberg (2005: 86). Both Arnaud (1998) and Smitterberg (2005) found that female 19th-century letter writers used more progressives than male ones. Moreover, in Smitterberg’s study the difference was shown to have increased over time. Gender and intimacy were the strongest factors in Arnaud’s study, with a positive correlation between the intimacy between the writer and the recipient of the correspondence and progressive frequencies.

In the 19th century, there was some prescriptive pressure against the progressive, but, as Smitterberg (2005) argues, this was largely restricted to the BE *being* progressive. Arnaud (1998) speculates that upper- and middle-class English women were affected by their Scottish and Irish servants’ dialects, which probably contained many progressives. Filppula (2008) argues that Celtic influence from Old English times is the main reason for English having typologically unusual features such as the *-ing* progressive. The factors indicating Celtic influence include the closest equivalent constructions to the English progressive being found in Celtic languages and the progressive being more frequent in ‘Celtic Englishes’.

3 Data and method

In-depth investigations of low-frequency phenomena such as subjective progressives require large data sets. The present material comes from the Corpus of Historical American English (COHA) (http://corpus.byu.edu/coha/; Davies 2012), which consists of 406 million words of written AmE from 1810–2009. The corpus covers four genres: fiction, non-fiction, magazines, and newspapers. About half of the material consists of fiction, while newspapers only begin to appear in the 1860s. The fiction sub-corpus largely contains novels and short stories, but there are some plays and, in later decades, movie scripts. The fiction material is therefore slightly different from the one used in previous studies (e.g., Leech et al. 2009; Smitterberg 2005), where drama and movie scripts are not included in fiction. The non-fiction sub-corpus covers a wide range of text types. It mainly comprises books about history, social science and technology. Some of the texts are clearly argumentative in nature, as *We can have Peace in the Holy Land* by Jimmy Carter, and yet others deal with self-help, such as *Emotional Freedom: Liberate yourself from Negative Emotions and Transform your Life* by Judith Orloff. From the 1990s, a number of journal articles from publications such as *Professional School Counseling* and *Journal of Environmental Health* are also included. The most recent decades in particular also include extracts from (auto)biographies. The magazine sub-corpus is sampled from a variety of sources, such as *Sports Illustrated*, *Good Housekeeping*, and *Time*. Finally, the newspaper data is sampled from a narrower range of publications, with half the material coming from *The New York Times*, which means that the results from this category must be viewed with some caution.

Some further limitations of the corpus should be mentioned. The 1810s contain only slightly more than one million words, and therefore the findings from this decade have a weaker empirical basis than the subsequent ones comprising 16–30 million words each. Moreover, some instances occurred twice in the corpus and some of the authors turned out to be non-American. However, the overall size of the corpus balances out most of these deficiencies.

The corpus was sampled in fifty-year intervals; 1810s, 1850s, 1900, 1950s and 2000s, and included all relevant tokens from these five decades. The restriction to five decades enables comparisons between the rarer adverbials and between genres without conflating different decades. The disadvantage of this approach is that any potential developments between the selected decades would go unnoticed. The advantages of comparing solid datasets from selected snapshots in time were nevertheless deemed to outweigh the disadvantages of overlooking possible variation in the data that may (or may not) be an effect of idiosyncratic sampling.

The material is restricted to three adverbials, *always*, *constantly* and *forever*, immediately followed by a progressive form of a verb (non-finite participial clauses such as *without always treading on people's toes* were excluded). BE *going to* forms creating future time reference were excluded, while motion verb instances of BE *going to* such as *he's always going to the bathroom* (COHA; Fiction; 2008) were included. The search string was *always*, *constantly* or *forever* immediately followed by **ing*. The five decades contained 2871 relevant instances which were classified for subject types and whether the verb was contracted. More than 60% of the tokens (1751) occurred in fiction. These were classified according to the gender of the author for the analysis of language change and gender in Section 4.4.

In spite of the restriction to only three adverbials, it is likely that the material covers a sizeable proportion of the subjective progressives, as will be shown in Section 4.1. In Smitterberg's (2005) data, 40% of the instances co-occurred with *always*, *constantly* or *forever*. Smitterberg's investigation was restricted to 19th-century British English, and, as will be seen below, the frequencies and distributions of the adverbials change over time. For example, some frequent adverbials in Smitterberg's (2005) study only rarely occur with the progressive in the 20th century.

Previous studies have employed different methods of calculating the frequencies of progressives. Aarts, Close, and Wallis (2013) argue that the proportional method, i.e. comparing frequencies of progressives with the frequencies of all "progressivisable" verbs, i.e. those verbs that could plausibly be turned into progressives (but were not), is preferable. However, they admit that the identification of "true alternants" with progressives "is, in part, subjective, and hence an approximation" (Aarts, Close, & Wallis 2013: 21). Smith and Leech (2013) show that identifying true alternants even in a superficially simple case as negative contractions cannot be carried out automatically (e.g., *-n't* contraction is not possible in instances such as *It's not mine* without additional changes). Therefore, they argue that using normalized frequencies (e.g., per million words) is both convenient and justifiable when studying highly complex cases of variation. It is difficult to determine which instances the subjective progressives with *always*, *constantly* and *forever* could be contrasted with. Accordingly, normalization per million words is the method employed in the present study. A further complication in a diachronic study is that it is difficult to determine when a verb is to be considered progressivisable or not at any given point in time.

In the present study, all instances of progressives co-occurring with *always*, *constantly* and *forever* were included among subjective progressives, since they are virtually always hyperbolic in nature and express some kind of subjective attitude.[1] This is seen in (3) where the hyperbolic nature of the construction is discussed.

(3) "Oh, sir," Allbright said, "that young gentleman *was always talking* about going away."
"Always?"
"Well, sir, maybe not exactly always. But enough to make it seem like always. That young gentleman was sort of uncontented around here."
(COHA; Fiction; 1959)

The examples were classified into two different groups according to the attitudes they express, either negative or positive/neutral. An example expressing negative attitudes such as annoyance, irony, irritation, sarcasm or disapproval is given in (4). In (5) there is one of the rare truly neutral instances, while (6) expresses a positive stance.

(4) Public officials *are forever doing* stupid things, but they don't step into hot tubs with naked women and cocaine unless they are driven to play Russian roulette with their careers. (COHA; Non-Fiction; 2003)
(5) Observations indicate that the different clusters of galaxies *are constantly moving* apart from each other. (COHA; Magazine; 1959)
(6) The best doctors I ever knew *were always trying* to make their patients live more simply, take more exercise, and give nature a chance; they never resorted to medicine until there was nothing else to do. (COHA; Fiction; 1901)

In (4), the writer is expressing his negative stance toward the typical behaviour of officials. The example is clearly hyperbolic, negative and subjective (since it is

[1] It is difficult to give an exact number of the "strictly neutral" instances since the examples form a continuum. A strict definition of "neutral" reduces the number to instances lacking all indications of negative or positive attitudes or the expression of intensification. The strictly neutral instances typically occur in scientific descriptions, as in (5) and (13), and include a few dozen examples.

literally not true that officials are forever doing this). In contrast, in (5)[2] there is little subjective attitude expressed. Finally, in (6) there is a positive attitude towards what the doctors were doing. In such cases there is a subjective expression of intensification.

The expression of attitudes was operationalized in the following way: if indications of negative attitudes such as annoyance, disapproval, irony or sarcasm were present, the examples were classified as expressing negative attitudes, otherwise they were classified as positive/neutral. It is often difficult to distinguish between neutral instances expressing intensification and instances expressing positive attitudes, and therefore these two categories were conflated. However, the distinction between negative on the one hand and positive/neutral on the other is often straightforward from the immediate context.

The distributions between negative and positive/neutral attitudes were investigated in order to explore whether there is ongoing subjectification. An increase in examples conveying negative attitudes would then be taken as an indication of the construction undergoing subjectification. This is an unusual way of operationalizing subjectification, but, as Traugott (2010: 56–60) points out, there are no defining criteria for subjectification that are valid across languages and constructions.

4 Results

The findings will be presented in the following order: Section 4.1 shows the distributions in the COHA sub-corpora and the frequencies for the individual adverbials. The overall frequencies of progressives in COHA are also discussed. Section 4.2 covers the influence of colloquialization on the spread of subjective progressives, and 4.3 explores the attitudes expressed. Finally, 4.4 deals with the influence of gender on language change.

4.1 The frequencies of subjective progressives and progressives in general across genres

Figure 1 presents the normalized frequency of BE + V-*ing* in COHA.

[2] Ljung (1980: 28–29) notes that although the progressive is typically associated with temporariness, it is also sometimes used with non-temporary events to focus on the middle phase of an event.

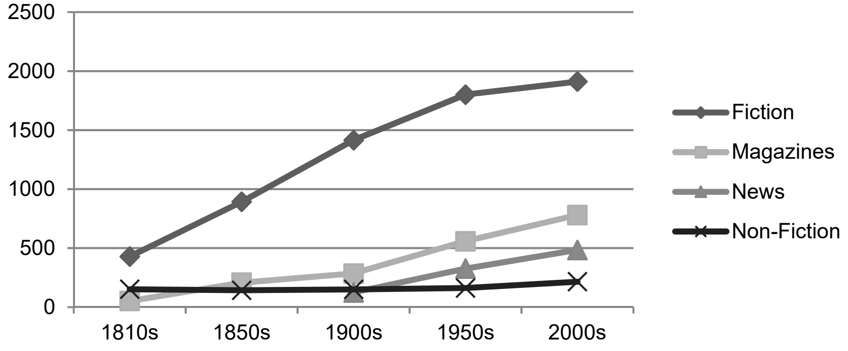


Fig. 1: BE + V-*ing* in COHA (normalized frequencies per million words).

Figure 1 shows that the progressive is becoming more frequent in all four genres, and that fiction contains the most progressives in COHA.[3] The frequency in fiction increases the most, but in the late 20th century the increase has been trailing off. This genre contains a lot of dialogue, and is therefore in some respects closer to speech than the other genres. However, this frequency is not solely due to progressives occurring in dialogue. Leech et al. (2009) note that in particular past progressives often occur in fictional narratives. The normalized frequencies in newspapers and magazines increase more slowly but due to their very low initial frequencies, progressives in these genres in fact increase proportionately more than in fiction, which means that there is a decreasing difference between fiction and magazines and newspapers, respectively. Progressives in non-fiction change the least, but still increase by 40% over the 200-year period. A likely explanation is that this genre is the least interpersonal in the expression of writer attitude and typically a 'slow' genre when it comes to adopting changes (for discussions of changes across genres, see, e.g., Kranich 2010a; Leech et al. 2009; Smith & Leech 2013).

Next, we turn to the overall frequencies of *always*, *constantly* and *forever* + progressive in COHA. Figure 2 shows that the subjective progressives more than double over 200 years. This is solely due to the increase in *always* + progressive, as was also found by Laitinen and Levin (2016) in the *Time* corpus. *Always* + pro-

[3] COHA does not contain separate sub-genres for trials, drama and letters, which means that it is difficult to make further comparisons with the CONCE where these genres produced the highest frequencies of progressives in 19th-century BrE (see Section 2).

gressive produces 49% of the instances in the 1850s, and this proportion increases to 80% in the 2000s. While *always* + progressive trebles in frequency, *forever* + progressive remains stable over time, and *constantly* + progressive decreases after peaking in the 1850s. Thus, the COHA material allows us to conclude that there is no wholesale increase in subjective progressives. Instead, this increase is lexically restricted.

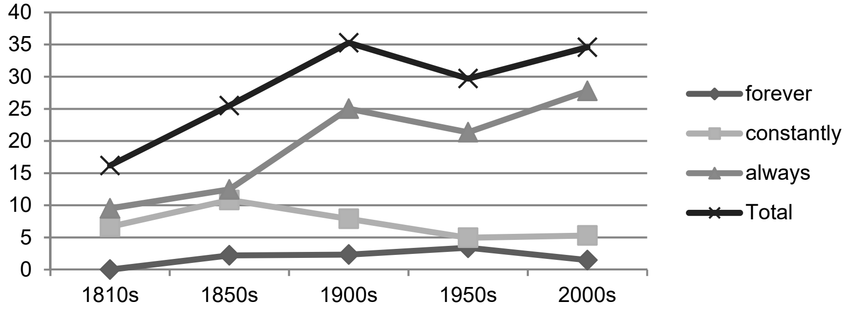


Fig. 2: *Always/constantly/forever* + progressive in COHA (per million words).

In Figure 2 there is a peak in the 1900s for *always*, but 12% (89 instances) of all tokens in this decade occur in a single text, *Three Lives* by Gertrude Stein. This shows that even with such a large corpus as COHA, individual texts may skew findings. Without this text, the total numbers for the 1900s are similar to those for the 1950s. Because this novel only affects the middle of the curve and not either of the end points, it was decided not to exclude this novel from the data.

The increase in subjective progressives cannot be explained with reference to changing frequencies of the adverbials *always*, *constantly* and *forever*, since their token frequencies have remained stable, at least since the mid-1800s. Furthermore, the inclusion of some of the other more common adverbials from Smitterberg's (2005) study would only marginally affect the results, and these effects would be virtually restricted to the 1800s. For instance, *continually* would only add 22 tokens (less than 1 per million words) to the overall number of 1,003 in the 2000s while this particular adverb was almost four times more frequent in the 1850s. *Perpetually*, which is also among Smitterberg's more frequent types, yields only five tokens in the 2000s (in comparison to 38 in the 1850s). Thus, the addition of more adverbials would probably make the increase of subjective progressives slightly less steep, but the overall findings would not radically change. The decrease for *continually* and *perpetually* supports the finding that subjective progressives are becoming specialized in that *always* is the adverbial that is taking over much of the territory.

Figures 1 and 2 show that both progressives overall and subjective progressives with *always/constantly/forever* are increasing. Different studies have found diverging results regarding subjective progressives in comparison to progressives in general. Mair and Hundt (1995) and Smith (2002) argue that the increasing use of subjective progressives has contributed greatly to the increase in the progressive. However, Smitterberg (2005) finds no change in the proportion of such progressives, and Kranich (2010a) finds a significant decrease in the proportion of subjective progressives in the ARCHER corpus between the 17th and 20th centuries. Figure 3 below shows that a similar pattern is seen in COHA.

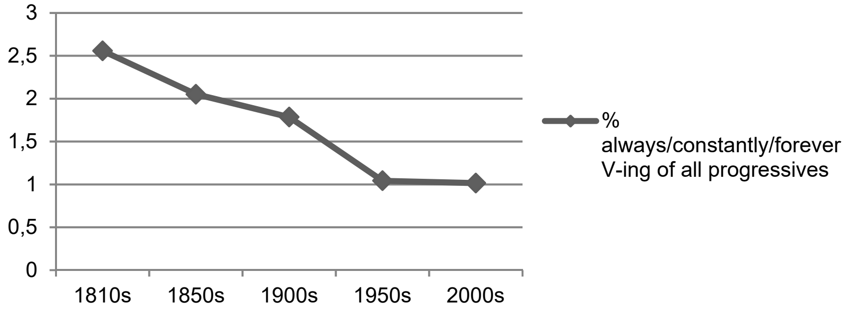


Fig. 3: Percentage of *always/constantly/forever* + progressive of all BE + V-*ing* in COHA.

Although subjective progressives are on the rise in COHA as regards their absolute frequency (see Figure 2), their proportion has been decreasing over the last 200 years due to the rapid increase of progressives overall (see Figure 1). Thus, subjective progressives only have a marginal influence on the overall increase of the progressive.

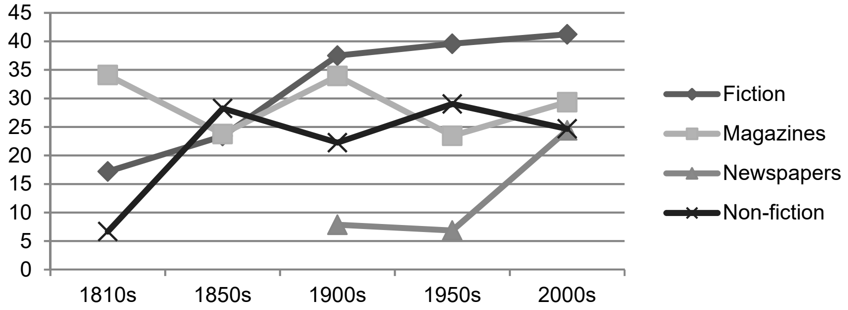


Fig. 4: *Always/constantly/forever* + progressive per genre in COHA (per million words).

The breakdown per genre presents a complex picture in Figure 4. There is an increase in subjective progressives in two genres only, fiction and newspapers. In contrast to Leech et al.'s (2009: 118–43) finding that progressives are spreading substantially in fiction, the increase in this genre has been slowing down in COHA in the 20th century, while the rate of change in newspapers has been increasing. The increase in fiction between the 1850s and 1900s and the increase in newspapers between the 1950s and 2000s were found to be significant ($p < 0.05$), using a log-likelihood test.[4] This change in newspapers is probably parallel to the steep increase in present active progressives in BrE newspapers in the late 1900s found by Smith and Leech (2013: 89). The deceleration of the increase in fiction may be caused by this genre slowly reaching a saturation point in the use of subjective progressives, since the proportion of dialogue cannot increase indefinitely. An example from fictional dialogue is provided in (7) where a character expresses a negative attitude towards her own behaviour.

(7) Oh, how awkward of me! I'm so sorry. *I'm always doing things like this*. (COHA; Fiction; 1956)

Subjective progressives are significantly ($p < 0.05$) more frequent in fiction than in the other 2000s sub-corpora,[5] but the fact that the frequencies in magazines, newspapers and non-fiction are very similar in the 2000s is perhaps less expected

[4] The UCREL Significance Test System (http://corpora.lancs.ac.uk/sigtest/) was used for all log-likelihood test. Fiction: 1850s 213 tokens; 1900s 450 tokens; newspapers 1950s 24; 2000s 100.

[5] Based on a log-likelihood test with 602 tokens in fiction, 224 in magazines, 100 in newspapers and 77 in non-fiction.

in view of, for instance, modal + progressive (e.g., *will be starting*) being more frequent in fiction and press than in general prose and learned texts in BrE from the Brown family of corpora (Smith & Leech 2013). Moreover, non-fiction and magazines have remained at similar levels for 150 years. However, these similarities can largely be explained by the composition of the corpus. Many instances in the non-fiction sub-corpus occur in books written in involved, argumentative style or in (auto)biographies and other text types close to fiction. For instance, the 1850s sub-corpus contains Harriet Beecher Stowe's *Sunny Memories of Foreign Lands*, which, as illustrated in (8), is, at least in parts, quite subjective and involved. In the example, the author expresses her displeasure at certain opinions expressed by others.

(8) The conversation now went on to Milton and Shakspeare. Macaulay made one remark that gentlemen *are always making*, and that is, that there is very little characteristic difference between Shakspeare's women. Well, there is no hope for that matter; so long as men are not women they will think so. (COHA; Non-Fiction; 1854)

In the 2000s, non-fiction contains extracts from highbrow to mass-market biographies such as *Melville: His World and Work*, *The Protest Singer: An Intimate Portrait of Pete Seeger* and *Juiced: Wild Times, Rampant 'Roids, Smash Hits and How Baseball Got Big*. (Auto)biographies have been found to be close to fiction, as noted both by linguists (e.g., Taavitsainen 1997) and literary scholars (e.g., Marcus 1994). The following extract from *Juiced* contains a number of first-person pronouns in baseball player Jose Canseco's description of his own childhood, and thus illustrates the sometimes highly personal and involved style in the sub-corpus. This relatedness to fiction means that subjective progressives are relatively frequent here as well.

(9) My parents and older sister, Teresa, were living in Regla in July 1964 when my mother gave birth to me and my twin brother, Osvaldo. People like to say that Ozzie and I were like pocket-sized atom bombs when we were babies, but my father says we were actually nice and quiet. *People were always fussing over us*. (COHA; Non-Fiction; 2005)

A comparison with the Longman Spoken American Corpus (Laitinen & Levin 2016) shows that the most advanced genre in the present study, fiction, is lagging behind AmE conversation from the 1990s. However, the difference between fiction and AmE conversation is small and non-significant[6], 41 vs 47 per million words. It is also noteworthy that the predominance of *always* + progressive is even stronger (89% of all tokens) in conversation than in the written genres in Laitinen and Levin's (2016) study. This suggests that the specialization may have progressed even further in speech.

4.2 Colloquialization

If writing in general is becoming colloquialized and subjective progressives are most frequently found in colloquial genres, then it is a reasonable hypothesis that these progressives increasingly often co-occur with other informal or involved features, such as contracted verb forms and first- and second-person pronoun subjects. Taavitsainen (1997) notes a connection between personal affect and such pronouns, and Römer (2005), Levin (2013) and Laitinen and Levin (2016) find correlations between such subjects with subjective progressives. However, Traugott (2010) argues that English first- and second-person pronoun subjects are not necessarily linked to increasing subjectivity, and suggests that negative attitudes typically correlate with third-person subjects.

Examples (10) and (11) illustrate first- and second-person pronoun subjects occurring with negative attitudes. In both examples, the verb forms are contracted, and both occur in fictional dialogue. They were classified as expressing negative attitudes due to the inherently negative semantics of the verb (*complaining*) or of the words or phrases in the immediate context (*wrong*; *repentant*).

(10) "*I'm always saying* the wrong thing," said the girl, in a repentant voice; (COHA; Fiction; 1904)

(11) Since you started your business, *you're constantly complaining* about money. (COHA; Fiction; 2009)

Figure 5 presents the percentages of first- and second-person pronoun subjects in the four COHA genres. The numbers for first- and second-person pronouns were conflated because the combined frequencies of these give a fairly good representation of the colloquialization of the texts. Until the 1950s, first- and second-

[6] Using a log-likelihood test ($p > 0.05$) and comparing with 236 tokens from 5.1 million words in LSAC (Laitinen & Levin 2016: 240).

person pronouns are about equally frequent but in the 2000s, first-person pronouns represent around two-thirds of all tokens.

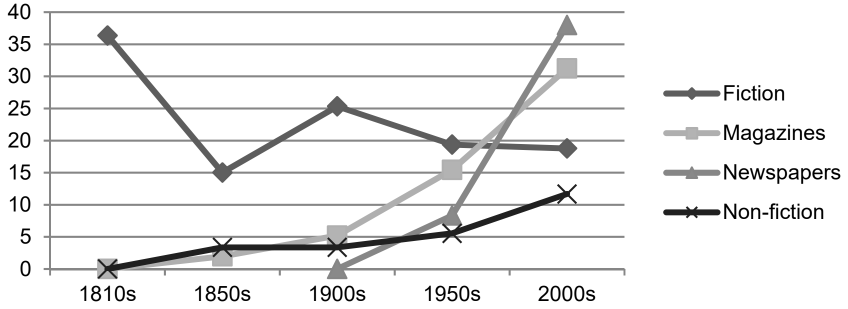


Fig. 5: Percentages of first- and second-person pronoun subjects with *always/constantly/forever* + progressive per genre in COHA.

Figure 5 shows that for most genres there is a steady increase in the use of first- and second-person subjects with subjective progressives. The increase is significant ($p < 0.05$) for magazines and newspapers between the 1950s and the 2000s, and for magazines between the 1900s and the 1950s.[7] The frequency of these subjects has stabilized in fiction, just like the frequency of subjective progressives in general as seen in Figure 4. It is likely that this genre already had a high proportion of fictional dialogue a century ago and therefore high proportions of first- and second-person pronoun subjects. In non-fiction the levels are slowly approaching those found in fiction, and, as argued above, this is an indication that the texts sampled in fiction and non-fiction share many similarities in the later decades. In the 2000s, there are significant differences between the genres with subjective progressives: fiction contains significantly fewer first- and second-person subjects than magazines and newspapers, and the latter two genres contain significantly more such subjects than non-fiction ($p < 0.05$).[8] There was also a

[7] Using a chi-square test; magazines: 1900s: 5% (9/173) (Phi coefficient = 0.1717), 1950s: 15% (21/136) and 2000s; 31% (70/224) (Phi coefficient = 0.1764); newspapers: 1950s: 8% (2/24) and 2000s: 38% (38/100) (Phi coefficient = 0.2507). The Phi coefficients indicate small correlations.

[8] Using a chi-square test; fiction 19% (113/602) (Phi coefficient for magazines 0.1336 and for newspapers 0.1636), magazines 31% (70/224), newspapers 38% (38/100) and non-fiction 12% (9/77) (Phi coefficient for magazines 0.1940 and for newspapers 0.2954). The Phi coefficients indicate small correlations.

notable, but non-significant,[9] difference overall between *always*, which is used with first- and second-person pronoun subjects in 25% of the instances, and *forever*, which is so in only 9% of the cases in the 2000s. A similar difference between the adverbs was found by Laitinen and Levin (2016) in *Time*. This is a further indication that the usage patterns of the three adverbs are drifting apart and may become specialized.

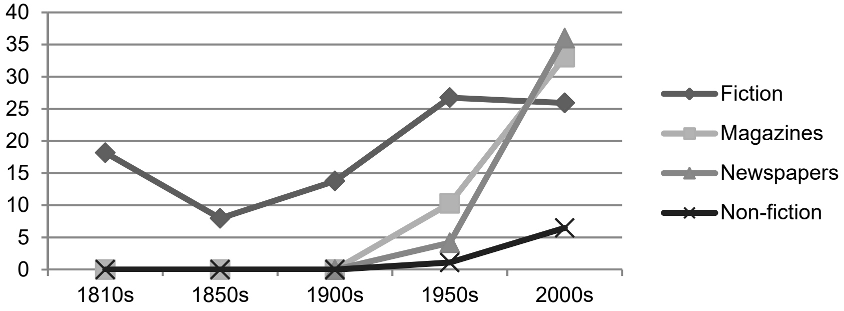


Fig. 6: Percentages of contracted verb forms with *always/constantly/forever* + progressive per genre in COHA.

Regarding contracted verb forms with *always*/*constantly*/*forever* + progressive, Laitinen and Levin (2016) found a significant increase in the last decades of the 20th century in *Time*. Figure 6 shows the diachronic developments of contracted verb forms in COHA and illustrates a similar pattern as Figure 5 above: towards the end of the 20th century, there are significant increases in the colloquial feature under investigation in magazines and newspapers ($p < 0.05$).[10] In non-fiction there is a smaller, non-significant rise. Once again, fiction differs from the other genres as the increase in contracted verb forms has levelled. Nevertheless, between the 1900s and the 1950s there is a significant increase ($p < 0.05$)[11] in con-

[9] Using a chi-square test: *always* (199/806); *forever* (4/43) (p-value = 0.03388; Phi coefficient 0.0791).

[10] The numbers for the 1950s and the 2000s were the following: magazines 15% (14/136) and 33% (74/224) (p-value = 0.000002; Phi coefficient = 0.2566, indicating a small correlation), and in newspapers 8% (1/24) and 36% (36/100) (p-value = 0.0049; Phi coefficient = 0.2749, indicating a small correlation).

[11] There was an increase from 14% (62/450) to 27% (127/475) contractions between the 1900s and 1950s (p-value = 0.0000016; Phi coefficient = 0.1606, indicating a small correlation).

tractions also in this genre. In the 2000s, there are significantly fewer contractions in non-fiction than in the other three genres ($p < 0.05$).[12] In the 2000s, there is also a significant difference (using a chi-square test $p < 0.05$) between *always* + progressive (31%) on the one hand, and *constantly* (15%) and *forever* (5%) on the other in the use of contractions.[13] A similar significant difference was also found by Laitinen and Levin (2016) in *Time*.

AmE writing habits have thus colloquialized, a change coinciding with the increase of subjective progressives. There has also been an increasing differentiation between genres, with magazines and newspapers colloquializing at a faster rate than non-fiction, while fiction has remained stable at least since the 1950s. It is also clear that *always* + progressive is not only the most frequent means of expressing subjective progressives, but is also becoming the more colloquial of the three.

4.3 Negative and positive/neutral attitudes

This section investigates shifting preferences for negative attitudes. Kranich (2010a) found a possible increase in negative attitudes across the 20th century, while there was no change over time in Laitinen and Levin's (2016) *Time* data. However, as mentioned in Section 2, Laitinen and Levin's (2016) findings suggest that there is an ongoing specialization in that *constantly* + progressive has a significantly lower proportion of negative attitudes than *always* and *forever*. Instead, Laitinen and Levin's (2016) *Time* material indicates that *constantly* is more connected with gradual change and inanimate subjects than with the expression of subjective attitudes, as illustrated in (12) to (14).

(12) I'm not quarreling. You*'re always quarreling and criticizing*. (COHA; Fiction; 1951)

(13) They found that stem cells in the brain *are constantly churning* out new neurons, particularly in the hippocampus. (COHA; Magazine; 2005)

[12] Using a chi-square test; the numbers were the following: fiction 26% (156/602), magazines 33% (74/224), newspapers 36% (36/100) and non-fiction 6% (5/77) (in comparison with fiction p-value = 0.00028; Phi coefficient = 0.1448; magazines p-value = 0.000010; Phi coefficient = 0.2632; newspapers p-value = 0.0000093; Phi coefficient = 0.3467).

[13] The numbers for the adverbs were the following: *always* 31% (246/806), *constantly* 15% (23/154) and *forever* 5% (2/43). P-value *always* vs. *constantly* = 0.00011; Phi coefficient = 0.1274; p-value *always* vs. *forever* = 0.00053; Phi coefficient = 0.1247.

(14) I'*m forever thinking* of ways to give design that extra kick, using unlikely sources. (COHA; Fiction; 2006)

Example (12) expresses a negative attitude towards a second person with the verbs *quarreling* and *criticizing*. In contrast, (13) is a rare 'neutral' example without clear positive or negative attitudes (as discussed in Section 3) included in the positive/neutral category. This sentence refers to an ongoing process without a human agent, and is therefore more easily perceived of in neutral terms. Finally, (14) expresses intensification of the activity described, as is probably partly also the case in (12) and (13).

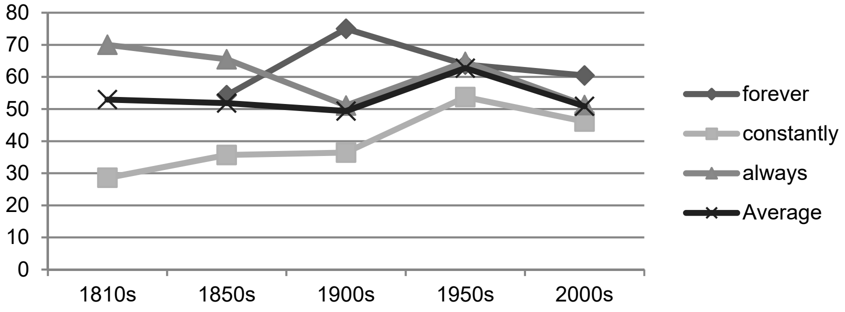


Fig. 7: Percentage of negative attitudes with *always/constantly/forever* + progressive in COHA.

The results for negative attitudes for each adverb are given in Figure 7. Overall, the attitudes remain fairly stable over time and the proportions of negative (53.6%; 1540 tokens) and positive/neutral (46.4%; 1331 tokens) are fairly equally distributed. There is, however, a significant decrease in negative attitudes between the 1950s and the 2000s.[14] The results are similar to those found by Laitinen and Levin (2016: 244) in that *constantly* + progressive has the lowest and *forever* + progressive highest proportion of negative attitudes. In the present study, however, the differences are not significant between the adverbs in the 2000s. The trend towards roughly equal proportions of negative attitudes for the three ad-

[14] 1950s: 455 negative; 270 positive/neutral; 2000s: 510 negative; 493 positive/neutral; p<0.001; Phi coefficient = 0.1184, which indicates a small correlation.

verbs therefore requires some explanation. With *always* + progressive there is a slight decrease in the negative attitudes over the two centuries, but this trend is due to some specific features of the material: there are only ten tokens from the 1810s, the drop in the 1900s is largely due to the novel *Three Lives* and the slight decrease remaining is largely connected to a tendency in newspapers and magazines in the 2000s to use *always* + progressive as a (positive) intensifying device, as in (15), rather than to express negative attitudes.

(15) She's never satisfied with her mark... Nicole *is always pushing* for that next level. (COHA; Newspapers; 2006).

The increase in negative attitudes with *constantly* is mainly caused by a shift in the largest sub-corpus, fiction, which has a larger proportion of examples expressing negative attitudes in the later decades. *Forever* + progressive remains rather stable over time but with a slight peak in the 1900s, which is due to an unusually large proportion of negatives in a number of texts in the fiction subcorpus.

As was noted for Figure 7, many of the fluctuations over the years with the individual adverbs are due to fluctuations in specific genres. Therefore, we now turn to the percentages of negative attitudes in each genre in Figure 8. Fiction remains quite stable over time, and it also remains different from the other genres by maintaining the highest proportion of negative attitudes. The difference between fiction and magazines and newspapers in the 2000s is significant ($p < 0.05$).[15]

[15] Fiction: 64% (327/602), magazines: 31% (70/224), newspapers: 24% (24/100) and non-fiction: 43% (33/77) (fiction vs. magazines p-value = 5.8×10^{-9}; Phi coefficient = 0.2053; fiction vs. newspapers p-value = 3.6×10^{-8}. Phi coefficient = 0.2119; in both instances the Phi coefficients indicate small correlations).

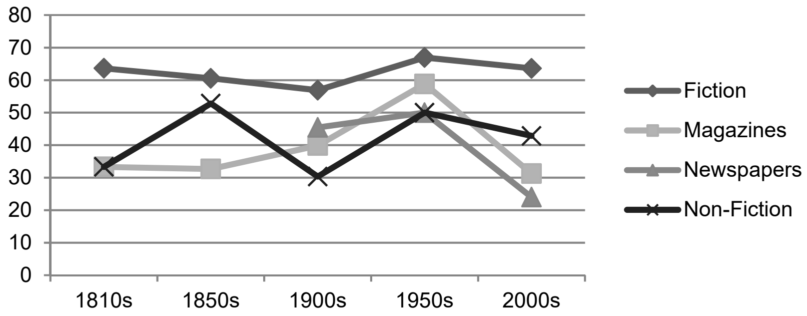


Fig. 8: Percentage of negative attitudes with *always/constantly/forever* + progressive in COHA per genre.

The reason for the consistently higher proportion of negative attitudes in fiction seems to be that characters express negative subjective attitudes in fictional dialogue, and, to a lesser extent, that the narrative also contains such attitudes. A typical example is given in (16), where the negative attitude is obvious from the subject, rather than the verb.

(16) "Where's Randy? He doesn't answer his phone."
"Dunno. Try him at the office. Little bastard*'s always working* late."
(COHA; Fiction; 2006)

In spite of some fluctuations, all genres show a decrease in negative attitudes between the 1950s and 2000s. This drop from 63% (455/725) negative attitudes in the 1950s to 51% (510/1003) in the 2000s is significant (p < 0.05) at the corpus level.[16] The decrease is also significant in two genres: in newspapers there is a significant decrease from 59% (80/136) to 31% (70/224), and in magazines from 50% (12/24) to 24% (24/100).[17] Arguably the strongest subjective meanings are expressed when negative attitudes are conveyed. Moreover, it is likely that this decrease in negative attitudes is linked to Kranich's (2010b) suggestion that subjective meanings tend to be more frequent with constructions in the early phases of grammaticalization, but that these attitudes are often lost in later stages.

[16] Using a chi-square test; p-value = 0.0000011; Phi coefficient = 0.1184, (small correlation).

[17] Using a chi-square test; for magazines p-value = 4.79×10^{-7}; Phi coefficient = 0.2712; for newspapers p-value = 0.023; Phi coefficient = 0.2263, in both cases indicating small correlations.

The decreasing proportions of negative attitudes in the 2000s are connected to the subjects involved. First-person subjects are becoming increasingly frequent and, as can be seen in Figure 9, the proportions of negative attitudes differ between the different subject types.

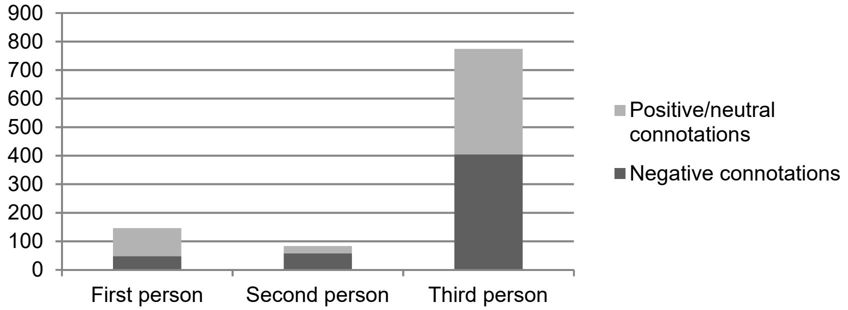


Fig. 9: Negative and positive/neutral attitudes with *always/constantly/forever* + progressive in COHA 2000s.

Figure 9 shows significant differences between the three persons.[18] A larger proportion of first-person subjects is connected with positive attitudes and a larger proportion of second-person pronoun subjects with negative attitudes. With first-person subjects, as in (14) above, the construction normally expresses positive attitudes or simply intensification. In contrast, (12) is a typical negative example with a second-person subject. Speakers or writers are thus more likely to refer to themselves rather than other people in positive contexts, while speakers apparently vent their frustration at the hearer's behaviour. A similar, but weaker tendency to that in the 2000s is found in the 1950s, where negative attitudes are significantly more frequent ($p < 0.05$) with second-person pronoun subjects than with the first and the third person, but with no significant difference between the

[18] First person 48 negative/98 positive/neutral; second person 58/25; third person 404/370. Using a chi-square test, the differences between first-person subjects and second- and third-person subjects are significant at $p<0.001$ (Phi coefficient in the former case 0.3567, indicating a medium correlation, and in the latter 0.1412, indicating a small correlation. The difference between second- and third-person subjects is significant at $p < 0.01$ (Phi coefficient = 0.1049, indicating a small correlation).

first and the third person.[19] Negative attitudes have decreased significantly ($p < 0.01$) with first-person subjects between the 1950s (53%) to the 2000s (33%),[20] which partly explains the decline in negative attitudes in Figure 8.

The distribution of subject types sheds further light on the findings presented in Figure 8. The number of first-person subjects is quite high in newspapers in the 2000s, while third-person subjects are frequent in non-fiction. Most of the first-person subjects in newspapers occur in direct quotations. The different distributions of subject types partly explain why newspapers contain so little negative attitudes in comparison with the non-fiction sub-corpus.[21] The subjective and interpersonal nature of these progressives is illustrated by 23% of the instances occurring with first- and second-person subjects in the 2000s. The proportion of first-person subjects increased notably, though not significantly from 8% in the 1950s (68/725) to 15% in the 2000s (146/1,003), while there was hardly any change for second-person subjects.[22]

As mentioned above, Traugott (2010) suggests that the frequency of third-person subjects increases when negative subjective attitudes increase with a construction. Judging from this, it could be expected that there would be a decrease in negative attitudes when the frequency of first-person subjects increases. The findings indicate that this is in fact the case when comparing the 1950s and the 2000s in COHA. Paradoxically, decreasing (negative) subjective attitudes are thus caused by slowly increasing first-person subjects, a change that is usually linked to increasing subjectivity. This change is also linked to Kranich's (2010b) idea that negative attitudes tend to decrease in the later stages of grammaticalization.

[19] Using chi-square tests; 2000s: Negative attitudes: first person 33% (48/146); second person 70% (58/83); third person 52% (404/774) (first vs. second person p-value = 1.4365×10^{-7}; Phi coefficient = 0.3567 (medium correlation); first vs. third person p-value = 0.000028; Phi coefficient = 0.1412 (small correlation); second vs. third person p-value = 0.0031; Phi coefficient = 0.1049 (small correlation). 1950s: first person 53% (36/68); second person 88% (46/52); third person 62% (373/605) (first vs. second person p-value = 0.000079; Phi coefficient = 0.3784 (medium correlation); second vs. third person p-value = 0.00021; Phi coefficient = 0.1506 (small correlation).

[20] Using a chi-square test. 1950s first person 53% (36/68); 2000s first person 33% (48/146) (p-value = 0.0081; Phi coefficient = 0.1913, indicating a small correlation).

[21] In newspapers from the 2000s, 26% (26/100) of the instances involved first-person subjects, while 88% (68/77) of the subjects in non-fiction were in the third person.

[22] 1950s: 7% (52/725); 2000s: 8% (83/1,003).

4.4 Language change and gender

Fiction is the only sub-corpus for which it was possible to identify authors' gender; the corpus website provides a list of all authors of the fictional works. Texts lacking authors, whose authorship remained unclear, were written by authors who could be identified as not being American or were written by multiple authors of both genders, were excluded. The excluded instances in the earlier decades are negligible[23] but account for 16.3% (98) of the 2000s material. The frequencies of progressives were compared to the numbers of words written by men and women. Since the material from the 1810s is so small, the study was limited to the 1850s onwards.

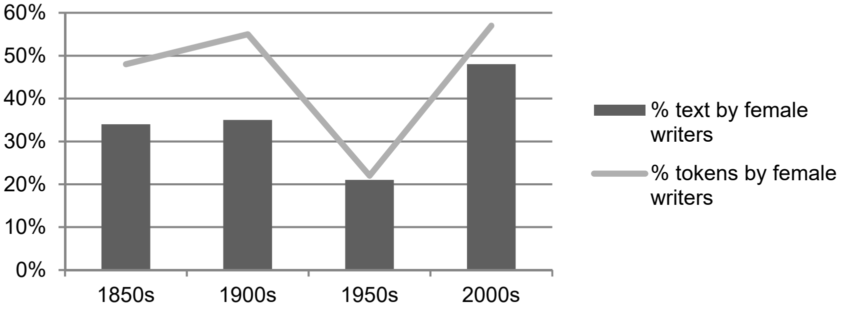


Fig. 10: Proportion of text and tokens of *always/constantly/forever* + progressive produced by female writers in the fiction sub-corpus of COHA, 1850s–2000s.

The results presented in Figure 10 show that, in three of the decades sampled, the 1850s, 1900s and 2000s, women produce significantly ($p < 0.05$) more progressives than men do in fiction, as determined by a log-likelihood test[24]. As fiction is the sub-corpus with the highest frequency of subjective progressives in COHA

[23] 7 instances in the 1850s (3.3%), 0 in the 1900s and 8 (1.7%) in the 1950s.

[24] In the 1850s, women produced 34% of the words, while producing 48% (99/206) tokens (LL 16.92). The Bayes Factor 0.93 indicates small correlation. 1900s: 35% of the text and 55% (247/450) of the tokens (LL: 75.53; the Bayes Factor 59.22 indicating very strong evidence against the null hypothesis of no correlation); 2000s, 48% of the words and 57% (286/504) of the tokens (LL: 15.98). In this decade, however, the Bayes Factor does not indicate a correlation.

(see Section 4.1), it is particularly striking that women are leading the increase in this particular genre. The results thus provide some support for the previous findings, both as regards language change in general and the progressive in particular. Women's greater propensity for using the progressive is thus established not only in BrE letter writing as found by Arnaud (1998) and Smitterberg (2005) but also in AmE fiction.

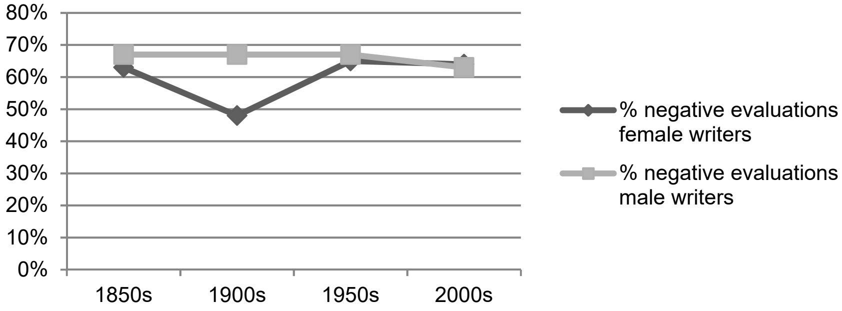


Fig. 11: Proportion of negative evaluations with *always/constantly/forever* + progressive produced by male and female writers in the fiction sub-corpus of COHA, 1850s–2000s.

As in Figure 8 above, there is no noticeable change in the proportion of negative attitudes in the fiction sub-corpus. Figure 11 presents the correlation between negative attitudes and gender of the authors in the fiction sub-corpus in COHA. It shows that there were similar overall distributions between the genders. The only notable difference is in the 1900s, where, once again, the novel *Three Lives* by Gertrude Stein affects the data with its idiosyncratic usage. There are, as noted above, an extraordinary number of instances of subjective progressives in that novel with very many of these expressing positive attitudes. The similar preferences in men's and women's usage in fiction are not surprising, since there are no significant indications of ongoing change in this genre. Thus, the COHA fiction data suggest that, although women are leading the way in the increase of *always/constantly/forever* + progressive, they are not different from men as regards the expression of negative subjective attitudes.

5 Conclusion

The findings from COHA largely support previous findings on the spread of the subjective progressive, but they also provide new insights into the changing patterns of usage in AmE. Subjective progressives have been increasing throughout the last two centuries, but there is no wholesale increase. In the 19th century, the spread was restricted to fiction, and in the late 20th century to newspapers. This increase can be attributed to the rise in *always* + progressive, while *constantly*, *forever* (and other adverbials) either remain stable or even decrease in frequency. The findings thus provide support for the ongoing specialization of these progressives as suggested by Laitinen and Levin (2016).

The spread of subjective progressives was found to be linked to the colloquialization of writing habits. However, the sampling of COHA to some extent affects the results. For example, the composition of the non-fiction sub-corpus covering everything from scientific journal articles to biographies and books on self-improvement makes it difficult to compare this sub-corpus with the others.

There is no indication of ongoing subjectification as expressed in increasing use of negative attitudes, but rather such attitudes are decreasing from the 1950s to the 2000s. The larger proportion of first-person subjects in these decades partly explains the more positive attitudes, but this change is possibly more strongly connected to Kranich's (2010b) suggestion that grammaticalizing constructions lose subjective shades of meaning in the later stages (but cf. Hübler's (1998: 91) discussions of the progressive as "a genuinely emotive language device"). It makes sense that constructions would tend to lose their subjective "edge" by increasingly frequent usage. A parallel case is the *get*-passive which, according to Mair (2006), has lost some of its adversative meaning during in the 20th century.

As could perhaps be expected from their typically involved style, conversation, as found by Laitinen and Levin (2016), and fiction in the present COHA material (which to a large extent consists of fictional dialogue) are genres where subjective attitudes are frequently expressed. The COHA data also show that women are leading the increase of *always/constantly/forever* + progressive in fiction, as has been found previously for progressives in letter writing (Arnaud 1998). However, there was no indication of women using more or fewer instances expressing negative attitudes than men do.

The present study has shown how a large-scale corpus can be used to study infrequent phenomena such as subjective progressives. Very large material is needed for the exploration of lexical phenomena, and COHA is large enough to provide insights into the changing patterns of the individual adverbials, and thus the ongoing specialization of subjective progressives.

Subjective progressives are increasing in frequency and this is partly connected to colloquialization (see e.g Laitinen & Levin 2016), while one of the other proposed main factors, subjectification, does not play a significant part (cf. Kranich 2010b). Subjective progressives are nevertheless not increasing faster than progressives in general, and contribute relatively little to the increase of progressives because of their overall low frequency.

Leech et al. (2009) suggest that the English progressive is increasing under its own momentum with frequency begetting increase. This is not the case with subjective progressives, which instead seem to become specialized to *always* + progressive (cf. Laitinen & Levin 2016) while increasing steadily, but perhaps not very quickly. Since there is no wholesale increase in frequency, subjective progressives appear to be an area which differs from the progressive in general. The full story of the English progressive therefore needs to be explored in detail in future studies of its various uses in different varieties, genres and time periods.

References

Aarts, B., Close, J., & Wallis, S. (2013). Choices over time: Methodological issues in investigating current change. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The Verb Phrase in English: Investigating Recent Language Change with Corpora* (pp. 14–45). Cambridge, UK: Cambridge University Press.

Arnaud, R. (1998). The development of the progressive in 19th century English: A quantitative study. *Language Variation and Change*, *10*(2), 123–152.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million-word Corpus of Historical American English, *Corpora, 7*(2), 121–157.

Filppula, M. (2008). The Celtic Hypothesis hasn't gone away: New perspectives on old debates. In M. Dossena, R. Dury, & M. Gotti (Eds.), *English Historical Linguistics 2006*. Volume III: Geo-historical Variation (pp. 153–170). Amsterdam, Netherlands: John Benjamins.

Finegan, E. (1995). Subjectivity and subjectivisation: An introduction. In D. Stein & S. Wright (Eds.), *Subjectivity and Subjectivisation* (pp. 1–15.). Cambridge, UK: Cambridge University Press.

Hübler, A. (1998). *The Expressivity of Grammar: Grammatical Devices Expressing Emotion across Time*. Berlin, Germany: de Gruyter.

Hundt, M. (2004). Animacy, agency and the spread of the progressive in eighteenth- and nineteenth-century English. *English Language and Linguistics*, *8*(1), 47–69.

Killie, K. (2004). Subjectivity and the English progressive. *English Language and Linguistics*, *8*(1), 25–46.

Kranich, S. (2007). Subjectification and the English progressive. The history of ALWAYS + progressive constructions. *York Papers in Linguistics* (Series 2) *8*, 120–137.

Kranich, S. (2008). Subjective progressives in seventeenth and eighteenth century English. Secondary grammaticalization as a process of objectification. In M. Gotti, M. Dossena, & R. Dury (Eds.), *English Historical Linguistics 2006. Selected Papers from the Fourteenth International Conference on English Historical Linguistics (ICEHL 14), Bergamo, 21–25 August 2006. Volume I: Syntax and Morphology* (pp. 241–256). Amsterdam, Netherlands: John Benjamins.

Kranich, S. (2010a). *The Progressive in Modern English. A Corpus-Based Study of Grammaticalization and Related Changes*. Amsterdam, Netherlands: Rodopi.

Kranich, S. (2010b). Grammaticalization, subjectification and objectification. In K. Stathi, E. Gehweiler, & E. König (Eds.), *Grammaticalization: Current Views and Issues* (pp. 101–121). Amsterdam, Netherlands: John Benjamins.

Kytö, M., Rudanko, J., & Smitterberg, E. (2000). Building a bridge between the present and the past: A corpus of 19th-century English. *ICAME Journal*, *24*, 85–97.

Labov, W. (2001). *Principles of Linguistic Change. Vol II: Social Factors*. Malden, MA: Blackwell.

Laitinen, M., & Levin, M. (2016). On the globalization of English: Observations of subjective progressives in present-day Englishes. In E. Seoane & C. Suárez-Gómez (Eds.), *World Englishes: New Theoretical and Methodological Considerations* (pp. 229–252). Amsterdam, Netherlands: John Benjamins.

Leech, G. (1987). *Meaning and the English Verb* (2nd ed.). London, UK: Longman.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English. A Grammatical Study*. Cambridge, UK: Cambridge University Press.

Levin, M. (2013). The progressive verb in modern American English. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The Verb Phrase in English: Investigating Recent Language Change with Corpora* (pp. 187–216). Cambridge, UK: Cambridge University Press.

Ljung, M. (1980). *Reflections on the English Progressive*. Gothenburg, Sweden: Acta Universitatis Gothoburgensis.

Mair, C. (2006). *Twentieth-Century English: History, Variation and Standardization*. Cambridge, UK: Cambridge University Press.

Mair, C., & Hundt, M. (1995). Why is the progressive becoming more frequent in English? A corpus-based investigation of language change in progress. *Zeitschrift für Anglistik und Amerikanistik*, *43*(2), 111–122.

Marcus, L. (1994). *Auto/biographical Discourses: Theory, Criticism, Practice*. Manchester, UK: Manchester University Press.

Núñez-Pertejo, P. (2004). *The Progressive in the History of English: With Special Reference to the Early Modern English Period. A Corpus-Based Study*. Munich, Germany: LINCOM.

Römer, U. (2005). *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts and Didactics*. Amsterdam, Netherlands: John Benjamins.

Smith, N. (2002). Ever moving on? Changes in the progressive in recent British English. In P. Peters, P. Collins, & A. Smith (Eds.), *New Frontiers in Corpus Linguistics* (pp. 317–330). Amsterdam, Netherlands: Rodopi.

Smith, N., & Leech G. (2013). Verb structures in twentieth-century British English. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The Verb Phrase in English: Investigating Recent Language Change with Corpora* (pp. 68–98). Cambridge, UK: Cambridge University Press.

Smitterberg, E. (2005). *The Progressive in 19th-century English. A Process of Integration*. Amsterdam, Netherlands: Rodopi.

Taavitsainen, I. (1997). Genre conventions: Personal affect in fiction and non-fiction in Early Modern English. In M. Rissanen, M. Kytö, & K. Heikkonen (Eds.), *English in Transition: Corpus-Based Studies in Linguistic Variation and Genre Styles* (pp. 185–266). Berlin, Germany: de Gruyter.

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In K. Davidse, L. Vandelanotte, & H. Cuyckens (Eds.), *Subjectification, Intersubjectification and Grammaticalization* (pp. 29–71). Berlin, Germany: de Gruyter.

Traugott, E. C., & Dasher, R. B. (2001). *Regularity in Semantic Change*. Cambridge, UK: Cambridge University Press.

Visser, F. Th. (1973). *An Historical Syntax of the English Language*. Leiden, Netherlands: Brill.

Wright, S. (1994). The mystery of the modal progressive. In D. Kastovsky (Ed.), *Studies in Early Modern English* (pp. 467–485). Berlin, Germany: de Gruyter.

Part III: **Learner contexts**

Tove Larsson

A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing

Abstract: The introductory *it* pattern, as in *It is important to consider the issue of learning outcomes*, is a versatile tool that has proved challenging for learners of English. Taking Quirk et al.'s (1985: 1392) seven syntactic types as the starting point, the present corpus-based study aims to map out the full inventory of this pattern in non-native-speaker and native-speaker student writing. Comparisons are made across native-speaker status, academic disciplines, and level of achievement (higher-graded papers vs. lower-graded papers). The material comprises student papers from three corpora: Advanced Learner English Corpus (ALEC), Michigan Corpus of Upper-level Student Papers (MICUSP) and the British Academic Written English (BAWE). The results show that while there are only small differences across native-speaker status, there are noteworthy differences across the academic disciplines. Furthermore, the students at a lower level of achievement show a preference for one syntactic type in particular. All in all, it seems that this pattern deserves a place among discipline-specific conventions taught to university students.

1 Introduction

The introductory *it* pattern, as in examples (1)–(4) is a multifaceted pattern of great importance to academic discourse. However, while instances of the pattern are commonly used by expert academic writers (Biber et al. 1999: 722), it is not unproblematic for learners to master (Hewings & Hewings 2002; Römer 2009).

(1) *It* is interesting *to note that age does not seem to be a factor* [...]. (ALEC_LING.131)[1]

(2) [...] *it* does not matter much *whether English is used* [...]. (ALEC_LING.105)

[1] The text ID for each corpus example is made up of three components: corpus (cf. Section 3.1), discipline and student code.

Tove Larsson, Uppsala University, tove.larsson@engelska.uu.se

https://doi.org/10.1515/9783110489071-012

(3) [...] *it* can be argued *that going into woods/forests is a metaphorical action for finding oneself.* (ALEC_LIT.021)

(4) *It* is an axiomatic truth *that 'Finnegans Wake' is a difficult text.* (ALEC_LIT.058)

As noted in previous studies, the pattern has a wide variety of uses, including functional and information-structural (e.g., Groom 2005; Hewings & Hewings 2002; Römer 2009). However, what a functional approach fails to register is that the diversity of the pattern also extends to its syntactic make-up. Indeed, most previous research on the introductory *it* pattern in academic discourse has focused only on instances of the pattern that include an adjective phrase, as in example (1) above, with the result that instances of the pattern such as the ones exemplified in (2)–(4) have remained unstudied.

The present study aims to complement previous research by approaching the introductory *it* pattern from a formal, syntactic perspective and thereby contribute to painting a more complete picture of the use of the pattern. Taking as its starting point Quirk et al.'s (1985: 1392) seven syntactic types (described in more detail in Section 3.2.2), the study presents the results of a syntactic analysis of the pattern as used by university students who are non-native speakers of English (NNS) or native speakers of English (NS). An investigation of not only what constitutes the full inventory of this pattern, but also the relative frequency of each syntactic type, will increase our understanding of the use of the introductory *it* pattern in student writing and enable identification of those syntactic types of the pattern that are likely to prove challenging for learners at different levels of achievement, thus facilitating more targeted teaching.

The overall frequency of the pattern has previously been investigated in both expert writing and apprentice student writing. Biber et al. (1999: 674, 722) looked at expert NS writing. Among other things they found occurrences of the introductory *it* pattern with an extraposed *to*-clause complementing an adjective to be slightly more frequent in academic prose than in news and fiction and considerably more frequent in academic prose than in conversation (Biber et al. 1999: 722). Expert and NS student writing were compared to NNS student writing in Römer's (2009) study, where the use of the introductory *it* pattern (of the kind *it* + *is* + (adverb) + adjective + *to/that*/OTHER-clause) was explored. Although some frequency differences were found between the groups, it was concluded that the differences found in the use of the introductory *it* pattern were likely to be due to students' general language proficiency and expertise in academic writing, rather than to their NS status (Römer 2009). However, Römer's study compared texts written by high-achieving NS students to texts written by NNS students whose production had been included in the corpus regardless of their level of achievement.

The introductory *it* pattern has furthermore been investigated across academic disciplines. Groom (2005) and Peacock (2011) looked at expert writing and found that the use of the pattern (of the kind *it* + linking verb + adjective + *to/that*-clause) seems to be discipline-specific. However, very few studies have been conducted on whether disciplinary differences manifest themselves in student writing. One exception is Thompson's (2009) study, which compared the use of *it* + BE/SEEM/APPEAR + adjective + *to/that*-clause in texts written by British students of history and engineering. Some differences across the disciplines were noted; the history students were, for example, found to use the pattern slightly more frequently overall. Possible differences across student levels were also investigated cross-sectionally, and an increase from the first to the third year of study was reported. The increase in use over time was concluded to be likely to be "an indication of a growing ability to express judgements within one's writing in an authoritative manner" (Thompson 2009: 79).

The material used for the present study allows for a more detailed investigation of the possible influence of NS status than was possible with the material used in Römer's (2009) study, as the use of the pattern both in high-achieving NS *and* in high-achieving NNS student writing will be compared. The study will furthermore investigate whether learners adhere to any potential discipline-specific uses of the pattern, similarly to the British students in Thompson's (2009) frequency-based study. The academic disciplines included for investigation in the present study are linguistics and literature. These two disciplines make for an interesting comparison, as they, in spite of their differences, are oftentimes placed in the same department in a European setting. Finally, the NNS material provides an opportunity to compare the use of the pattern across level of achievement for the learners (higher-graded papers vs. lower-graded papers). This has, to the best of my knowledge, not previously been done. In sum, the present comparative and corpus-based study maps out the frequency of use of the syntactic types of the pattern through comparisons across (i) NS status, (ii) academic disciplines, and (iii) level of achievement.

2 Defining the introductory *it* pattern

The introductory *it* pattern has, in previous research, commonly been referred to as '(subject) extraposition'. 'Extraposition' is defined by Quirk et al. (1985: 1391) as "[p]ostponement which involves the replacement of the postponed element by a substitute form”. It is primarily subordinate nominal clauses that can be extraposed (Quirk et al. 1985: 1391). While extraposition can operate on both clausal subjects and clausal objects, the present study only includes investigation of the more frequent of these, namely extraposition of clausal subjects (cf. Quirk et al. 1985: 1391). A sentence with an extraposed subject is shown in (5).

(5) *It* is important *to look at foregrounding*. (ALEC_LING.074)

According to Quirk et al. (1985: 1391), sentences such as this one are derived from sentences with “more orthodox ordering”, i.e. those which have a non-extraposed clausal subject (e.g. *To look at foregrounding* is important). The clausal subject of the sentence is analysed as having been moved and the subject position filled by the introductory pronoun *it*, resulting in a sentence that has two subjects (Quirk et al. 1985: 1391). Criticism has, however, been voiced against such an analysis, especially since the extraposed version is considerably more frequent (e.g. Mair 1990; Mindt 2011; Mukherjee 2006). Another reason for not viewing extraposition as a result of derivation is its failure to provide satisfactory explanations for instances of the pattern for which extraposition is obligatory, as in example (6), and for which there is no equivalent non-extraposed version (cf. the constructed example in (7)).

(6) *It* seems *that they have already crossed that line*. (ALEC_LIT.111)
(7) **That they have already crossed that line* seems.

In light of these points of criticism, the term ‘extraposition’ will not be used in the present study; instead, the term ‘introductory *it* pattern’ will be used (cf., e.g., Francis, Manning, & Hunston 1998; Groom 2005; Larsson 2016; Larsson 2017). The introductory *it* pattern is here defined as a pattern which contains two subjects, one of which is realized by introductory *it* (which does not have anaphoric reference) and the other by a nominal clause.

Many previous studies have restricted their analysis to instances of the introductory *it* pattern in which the clausal subject is realized by *to*-infinitive clauses and *that*-clauses (this is the case, for example, in Groom 2005 and Hewings & Hewings 2002). In this study, however, all of the following nominal clausal subjects have been included: *that*-clauses (8), *to*-infinitive clauses (9) (including *for/to*, as in (10)), *-ing* clauses (11), and *wh*-clauses (12).

(8) *It* appears *that her independence is a contributing factor*. (ALEC_LIT.118)
(9) [...] *it* has not been considered necessary *to make reference to the proposed underlying principles of UG*. (ALEC_LING.020)

(10) *It* becomes impossible *for anyone else than Zeus to master the task.* (ALEC_LING.083)
(11) *It* is great fun *learning English.* (ALEC_LING.003)
(12) [...] *it* has yet to be explained *why the degree modifier 'equally' behaves in largely the same way.* (ALEC_LING.083)

Instances that are not covered by the definition, such as *it*-clefts (13), referring *it* (14) and 'prop' *it* (15) were, however, excluded. Tokens which include an adverbial clause (16) rather than a nominal clause were also excluded, in accordance with the definition given above.

(13) *It* is during this process that the surprise effect is created. (ALEC_LING.024)
(14) Crystal [...] uses the term 'Netspeak' when addressing the Internet language. *It* [Netspeak] exhibits many features that complicate a traditional understanding of language. (ALEC_LING.033)
(15) On a car drive to Heinrich's school, Jack asks Heinrich if *it* is raining right now. (ALEC_LIT.037)
(16) *It* takes some time before she writes them a letter. (ALEC_LIT.038)

3 Design of the study

In this next section, the design of the study will be presented in more detail; the data and material used will be presented in Section 3.1, followed by the method in Section 3.2.

3.1 Data and material

The material used for the present study comprises subsets from three large corpora of university student writing: the Advanced Learner English Corpus (ALEC), the Michigan Corpus of Upper-level Student Papers (MICUSP) and the British Academic Written English (BAWE) corpus. The NNS corpus and the NS reference corpus were carefully sampled to ensure comparability to the greatest extent possible. Five factors of potential importance were controlled for in each corpus: the students' first language (L1), level of achievement, academic discipline, text type and contribution (i.e. the number of words each student has contributed).

ALEC includes mainly NNS writing by students whose L1 is Swedish. The NNS corpus used in the present study consists exclusively of L1 Swedish papers (approximately 590,000 words). The NS reference corpus comprises only texts by L1 English students from all three corpora (approximately 260,000 words). While the ALEC NS students' texts represent a number of different regional varieties of English, the BAWE students' texts represent British English (BrE) and the MICUSP students' texts represent American English (AmE). Since a small-scale pilot study indicated that the use of the introductory *it* pattern might differ between British and American English, it was considered important to have both varieties represented in the corpus. In total, approximately 840,000 words (135 texts) of student writing are included in the present study, which yielded a total of 1,710 valid tokens of the introductory *it* pattern. An overview of the subcorpora can be found in Table 1.

Tab. 1: Number of words per subcorpus.

Subcorpus	Number of words
BAWE (NS)	94,345
MICUSP (NS)	121,147
ALEC A+B (NS)	39,786
NS total	255,278
ALEC A+B (NNS)	361,965
ALEC C+D+E (NNS)	225,864
NNS total	587,829
Total	843,107

Only higher-graded papers are included in BAWE and MICUSP, with the aim of representing successful student writing (Heuboeck, Holmes, & Nesi 2008; Römer & O'Donnell 2011). Since students at all levels of achievement (grades A through E, all of which are passing grades) are included in ALEC, the NNS data was subdivided into two subcorpora; only the higher-graded NNS papers (that were awarded an A or B) were compared to the NS data. The lower-graded papers (grades C, D and E) make up the NNS subcorpus to be compared to the NNS subcorpus of higher-graded papers for the comparison across level of achievement. Only higher-graded papers were included in the NS subset of ALEC. The terms 'high-achieving students' and 'low-achieving students' are used to refer to the students who were awarded a higher grade (an A or B) and the students who received a lower grade (a C, D or E) respectively. While there are qualitative differences between the texts in the two categories, no general tests of the students' level of English proficiency have been carried out; therefore, no claims about the students' level of proficiency can be made in the present study. 'High-achieving' and 'low-achieving' are thus to be understood merely as descriptive labels (synonymous to 'higher-graded' and 'lower-graded' respectively), used in order to be able to refer to and compare the use of the pattern across the two groups of NNS students.

Furthermore, the comparison across academic disciplines was carried out between the two disciplines represented in ALEC, namely linguistics and literature; for this reason, only texts written by students of linguistics or literature were collected from BAWE and MICUSP. With regard to text types, since ALEC includes only papers reporting on research, text types of close resemblance to such texts were selected from BAWE and MICUSP. For linguistics, texts classified as "essays" or "research reports" were selected from BAWE (cf. Heuboeck, Holmes, & Nesi 2008), and texts categorized as "research papers", "argumentative essays" or "proposals" were selected from MICUSP (cf. Römer & O'Donnell 2011). For literature, only texts classified as "essays" were selected from BAWE (cf. Heuboeck, Holmes, & Nesi 2008) and "argumentative essays" and "critique/evaluations" were selected from MICUSP (cf. Römer & O'Donnell 2011). Despite the fact that the corpora are made up of seemingly different text types, all the texts included have a common core of being topic-based or thesis-driven research papers. For BAWE and MICUSP, it was the students who contributed their text who assigned it a category; all the texts that are included in the present study have therefore been checked manually for relevance.

The NS reference corpus and the NNS corpora in the present study were sampled to comprise texts written by students who are in their third or fourth year of studies on average. Due to the Swedish system for courses at undergraduate level, the third-year students whose texts are included in ALEC are, on average, in their third year of university studies in total, with all their coursework taken into account. The students have typically taken three terms (equivalent to 1.5 years) of English linguistics or English literature, as that is what is required to earn a first-cycle degree in English.

In order to account for the fact that the texts differ in size across BAWE, MICUSP and ALEC, upper and lower cut-off points were introduced with the purpose of excluding outliers and bringing the means for the number of words contributed per person closer across the corpora. For this reason, the students have contributed between 2,000 and 15,000 words, with the mean length being approximately 6,000 words (somewhat lower for the NS reference corpus and somewhat higher for the NNS corpus). If a student had contributed more than one text, resulting in the word count amounting to more than 15,000 words, the statistical software R (R Core Team 2015) was used to generate a random sample of that student's texts to be excluded, to keep the total word count to less than 15,000 words per student; only full texts were included.

Being very selective with the purpose of ensuring comparability to as large an extent as possible does, however, come at a price: about half the linguistics and literature texts that could have been included in the NS reference corpus did not meet the criteria and were therefore excluded. Similarly, approximately one third of the NNS texts did not meet the criteria (mainly due to the contribution criterion). Thus, while this procedure increased the degree of comparability across the corpora, it caused the size of the corpora to decrease considerably. Also, in spite of the careful sampling procedure, there are still certain unavoidable differences between the NNS corpus and the NS reference corpus, mainly pertaining to the length of the texts, the students' background knowledge of their subject and the amount of feedback the students have been given on their assignments by the instructors, which should be kept in mind when looking at the results.

Finally, although somewhat controversial in some subfields of linguistics, such as English as a Lingua Franca (cf., e.g. Jenkins 2006), one assumption that underpins the present study is that the NS corpus is seen as the reference corpus to which the NNS students' texts are compared. The main difference between the view adopted in this study and that of ELF lies in what is seen as the goal for learning English in an English as a Foreign Language context. In this study, along with other studies situated within the subfield of learner corpus research (cf., e.g. Granger 2002), the applied aim is for the results to be used for teaching; a point of reference is thereby required, which, in this case, is NS student use. A strong version of this view is presented by Granger (2002: 13) who states that

> [i]f learner corpus research has some applied aim, the comparison with native speaker data is essential since the aim of all foreign language teaching is to improve learners' proficiency, which in essence means bringing it closer to some NS norm(s).

In ELF studies, successful communication is seen as more important than speaking "correct" English; "ELF is thus a question, not of orientation to the norms of a particular group of English speakers, but of mutual negotiation involving efforts and adjustments from all parties" (Jenkins 2009: 201).

3.2 Method

In the subsequent subsections, the method will be described. The retrieval and analysis of the data will be addressed in Section 3.2.1, and the syntactic classification will be presented in Section 3.2.2.

3.2.1 Data retrieval and data analysis

While many previous studies have used search strings such as *it* + linking verb + adjective + *to*/*that*-clause, the present study aimed to achieve full coverage of the instances of the introductory *it* pattern in the data. For this purpose, WordSmith Tools Version 6 (Scott 2012) was used to find all instances of *it* in the corpora, excluding those that were part of quotations. Excluding all instances of the introductory *it* pattern that occurred as part of quoted material restricted the study to the investigation only of instances of the introductory *it* pattern produced by the students themselves. All hits were subsequently analysed manually in order to exclude invalid tokens, such as *it*-clefts and referring *it* (cf. Section 2).

In addition to the invalid tokens mentioned above, another category of excluded tokens (albeit a very small one) deserves mention. If I identified an instance of the pattern in the NNS data as infelicitous and the token was unattested in the multi-million-word NS corpora Corpus of Contemporary American English (COCA) (Davies 2008) and the British National Corpus (BNC), the token was excluded; the search strings used included all forms of the lemma and allowed for optional intervening adverbs. Only a handful of such tokens were omitted. Examples of this very marginal category are provided in (17) and twice in (18); the problematic lexical items are underlined.

(17) The students did not ask any questions. Maybe because it was very difficult but *it* is more probably *that the teacher was very good in explaining and therefore easy to understand*. (ALEC_LING.116)

(18) *It* was general during this time *to symbolically "strip the black Africans" of language*, just as *it* was frequent *to exclude them from setting*. (ALEC_LIT.055)

Although a corrected, felicitous version of each pattern could, in many cases, be arrived at by guessing (*probably*, *general* and *frequent* could potentially be changed to *probable*, *general practice* and *common* respectively), the infelicitous lexical item used inside the pattern was still thought to impede the transparency of the expression enough to justify exclusion, as it was considered important not to let the classification be guided by guesswork. This was especially important for the syntactic classification (cf. Section 3.2.2), since tokens such as the one exemplified in (17) would have been classified as belonging to different syntactic types depending on whether the current or the "corrected" version formed the basis for the categorization (SVA and SVC respectively). Only seven such tokens were found and excluded.

The manual screening allowed for the inclusion of realizations of the introductory *it* pattern that would be difficult to find through automatic retrieval using search strings, such as tokens with inverted word order (19) and tokens in which the complementizer *that* is omitted (20). Although more time-consuming than fully automatic retrieval, this approach moreover has a clear advantage in that it allows for inclusion of instances of the introductory *it* pattern which do not include adjectives, as exemplified in (20) and (21).

(19) Is *it* possible *to distinguish a primary metaphor from a complex metaphor*? (ALEC_LING.030)

(20) *It* seems *they are still on the threshold*. (ALEC_LIT.017)

(21) *It* is not a coincidence *that Grendel's mother takes Aeschere* [...]. (MICUSP_LIT.012.1)

A syntactic analysis was subsequently carried out, and all instances of the introductory *it* pattern were classified into syntactic types (based on Quirk et al. 1985: 1392), as explained in Section 3.2.2 below. Since the subcorpora differ in size, all frequencies were normalized. Using R, a log-rate generalized linear model (GLM) was fitted onto the results in order to test the differences for statistical significance and to investigate whether there were interacting predictors. Regression analyses of this kind take into account the fact that the size of all the subcorpora differed considerably (Powers & Xie 2000). R was furthermore used to perform Chi-square tests, as well as to calculate the frequency of occurrence of each syntactic type per text and, subsequently, to carry out a Kruskal-Wallis rank sum test on the medians of these frequencies in order to test the differences for statistical significance. Moreover, the study will not only present the results for the total frequencies across the different groups as a whole, but also, when relevant, for the dispersion of the frequencies across the individual texts.

3.2.2 Syntactic classification of the introductory *it* pattern

As mentioned earlier, the syntactic types used for the classification are based on Quirk et al. (1985: 1392). Two subcategories have been added to the first and the seventh categories, as is shown in Table 2; however, although present in the classification, the subcategories 7a and 7b will be conflated in the present study due to the infrequency of the $\text{SV}_{\text{pass}}\text{C}$ type. Only two tokens did not match any of these syntactic types; these were classified as OTHER. The study will not include investigation of what type of clause serves as the clausal subject, as not all syntactic types admit variation; the third syntactic type, SV, does not, for example, allow for *-ing*-clause complementation.

Tab. 2: Overview of the syntactic types of the introductory *it* pattern used for classification (based on Quirk et al. 1985: 1392).

Syntactic type	Corpus example
1. SVC: Subject + Verb + Complement	
1a: Complement realized by Adj. Phrase	It is important to look at foregrounding. (ALEC_LING.074)
1b: Complement realized by Noun Phrase	It is no surprise that this discourse is used as a rhetorical device. (ALEC_LING.112)
2. SVA: Subject + Verb+ (obligatory) Adverbial	It has been beyond the scope of this study to look at all the evaluative examples. (ALEC_LING.049)
3. SV: Subject + Verb	It seems that the city defends itself. (ALEC_LIT.067)
4. SVO: Subject + Verb + Object	It would involve considerable manual intervention to sift out all the false hits. (ALEC_LING.032)
5. SVOC: Subject + Verb + Object + Complement	It makes her happy to hear them quarrel. (ALEC_LIT.038)
6. SV_{pass}: Subject + passive Verb	It is shown how this metaphor incorporates a metonymic mapping. (ALEC_LING.130)
7. $\text{SV}_{\text{pass}}\text{C}$: Subject + passive Verb + Complement	
7a: Complement realized by Adj. Phrase	It was found necessary to broaden the existing definition. (ALEC_LING.112)
7b: Complement realized by Noun Phrase	It must be considered a sociolinguistic fact that the Swedish language is under the influence of English. (ALEC_LING.049)

Although generally relatively straightforward, the process of classifying the tokens was not entirely unproblematic. For example, instances which contain BE + a past participle and which seem, at first glance, to belong to the SV_{pass} category might instead potentially be classified as SVC, depending on whether the past participle is verbal or adjectival. A similar problem could potentially occur for tokens belonging to the $SV_{pass}C$ type; no such problematic cases were, however, encountered in this material. Following Quirk et al. (1985: 167–171), a distinction was made between instances that meet only the formal criteria (i.e. BE followed by a past participle) and instances that fulfil both the formal and the functional criteria. Quirk et al. (1985: 167) state that instances which meet the formal criteria can be placed on a scale ranging from 'central passives' to adjectival complements. Based on this 'passive gradient' (Quirk et al. 1985: 167–171), instances of the introductory *it* pattern were categorized as SV_{pass} if they could be paraphrased into an active sentence (albeit not in the form of an introductory *it* pattern) and if they did not meet the criteria for 'semi-passives' or 'pseudo-passives' (i.e. if they cannot, for example, be coordinated with an adjective, modified by degree adverbs and/or be placed after other copular verbs such as SEEM). One such instance of a central passive is shown in (22). Instances that met the criteria for semi or pseudo-passives were categorized as SVC (23); these were, however, comparatively infrequent.

(22) In sum, *it* can be concluded *that the VG-level texts score the highest results* [...]. (ALEC _LING.015)
(23) [...] *it* was generally accepted *that decent men and women of the period recognised sexual "honesty"* [...]. (ALEC_LIT.100)

Furthermore, while many instances of the introductory *it* pattern can be considered idioms or fixed expressions due to their relative lack of formal variability, all instances have been analysed as syntactic strings rather than as chunks for the purpose of the present study. Fixed expressions such as *It is no wonder that*... and *It goes without saying that*... are thus classified as SVC and SVA respectively.

4 Results and discussion

As stated in Section 1, the focus of the present study is on investigating how the frequencies of the syntactic types map out across academic disciplines, NS status, and level of achievement. Section 4.1 provides an overview of the results in terms of the overall frequency of occurrence of the introductory *it* pattern across the different subcorpora (4.1.1) and across the syntactic types (4.1.2). The results for each of the three comparisons are presented and discussed in the three subsequent Sections (4.2–4.4).

4.1 Overview

The overall frequencies of the introductory *it* pattern will be presented in Section 4.1.1, followed by a breakdown of the frequencies per syntactic type in Section 4.1.2.

4.1.1 Overall frequency of occurrence of the introductory *it* pattern

As mentioned in Section 3.1, a total of 1,710 instances of the introductory *it* pattern were included in the analysis; Table 3 shows the overall frequencies per subcorpus.

Tab. 3: Overall frequency of the introductory *it* pattern across the subcorpora.

Subcorpus	**Literature**		**Linguistics**		**TOTAL**	
	Raw freq.	Per 100,000 words	Raw freq.	Per 100,000 words	Raw freq.	Per 100,000 words
ALEC NNS A+B	283	146	402	239	685	189
ALEC NNS C+D+E	215	175	295	286	510	226
NNS total	498	157	697	257	1,195	203
BAWE (BrE)	83	146	115	307	198	210
MICUSP (AmE)	101	139	129	267	230	190
ALEC NS A+B	16	200	71	224	87	219
NS total	200	145	315	268	515	202
TOTAL	698	154	1,012	260	1,710	203

While slight differences can be noted when comparing the higher-graded NNS papers to the lower-graded NNS papers (189 and 226 occurrences per 100,000 words respectively) and the BrE texts to the AmE texts (210 and 190 occurrences per 100,000 words respectively), the main differences appear to lie in the discipline comparison.

In order to explore these differences further, a log-rate GLM was fitted onto the data, as shown in Table 4, to investigate the use of the pattern across the corpora. The model is not a ‘minimal model’, i.e. it does not include only the significant predictors, as the purpose of the model is to present a comparative overview of the different predictors included in the study. The model as such does not have very much explanatory power. Further, in order to provide an overview, two predictors – NS status (the NS papers vs. the higher-graded NNS papers) and level of achievement (the higher-graded NNS papers vs. the lower-graded NNS papers) – have here been conflated under the heading 'student group'.

Tab. 4: The output of a log-rate GLM of the overall frequency of the introductory *it* pattern.

Predictor	Estimate	Std. Error	z	p	Significance level
Intercept	-5.96	0.14	-43.64	<0.001	***
Student group: NNS A+B					
vs. NS total	-0.02	0.12	-0.17	0.869	
vs. NNS C+D+E	0.18	0.06	3.09	<0.01	**
Discipline: Linguistics					
vs. Literature	-0.53	0.05	-10.75	<0.001	***
Variety: AmE					
vs. BrE	0.10	0.01	1.05	0.296	
vs. Other	-0.06	0.13	-0.45	0.652	

Null deviance: 133.925 on 9 degrees of freedom
Residual deviance: 5.516 on 4 degrees of freedom
AIC: 84.29

As can be seen from the column of *p*-values (also illustrated in the significance-level column), there is no statistically significant difference in terms of overall frequencies of the introductory *it* pattern between the higher-graded NNS papers and the NS papers (*p*>0.05). There is, however, a statistically significant difference across level of achievement for the NNS students (*p*<0.01); the positive *z*-value indicates that the pattern is more frequently used in the lower-graded NNS papers than in the higher-graded NNS papers, which is consistent with Table 3. Interestingly, it is also clear that the literature texts include significantly fewer instances of the introductory *it* pattern than the linguistics texts across the corpora (*p*<0.001; *z*: -10.75). These differences and similarities between the groups will be further explored in Sections 4.2–4.4. There was, however, no statistically significant difference between BrE and AmE (*p*>0.05); the results for these two subsets will therefore not be presented separately, but instead as one NS corpus, along with the NS data from ALEC.

In what follows, we will go beyond the overall frequencies reported here and instead investigate the use of the pattern in more detail, by investigating the use of the syntactic types.

4.1.2 Overall frequency of occurrence per syntactic type

When it comes to overall frequencies across the syntactic types, considerable differences can be noted, as shown in Figure 1.

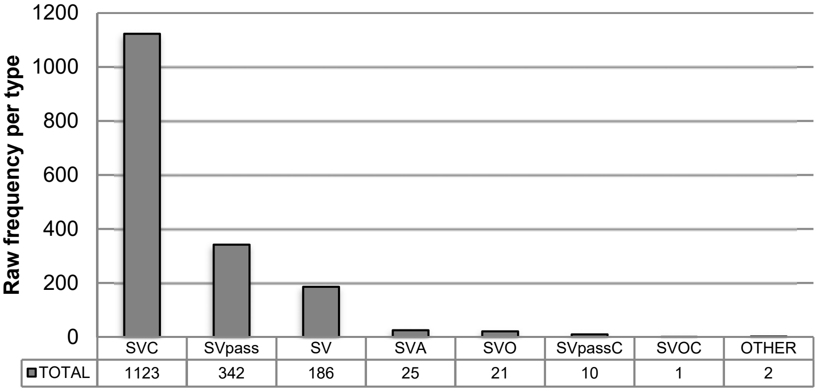


Fig. 1: Absolute (raw) frequencies per syntactic type, overall.

By far the most frequent syntactic types are SVC (as in example (24)), SV_{pass} (25) and SV (26); together these make up 97 percent (1,651/1,710) of all occurrences of the introductory *it* pattern. The distribution bears some resemblance to a Zipfian distribution (Zipf 1935) in which the frequency of items decreases as a function of their frequency rank; in practice, this means that a very limited number of items occur very frequently and the rest of the items occur very infrequently.

(24) Therefore, *it* is evident *that there is need for an investigation about culture and English teaching* [...]. (ALEC_LING.041)

(25) *It* could be argued then *that Esther's 'madness' lies less in subjugation by the state* [...]. (BAWE_LIT.3006l)

(26) [...] *it* seems *that most of his dealings with women end poorly* [...]. (MICUSP_LIT.049.2)

Due to the considerable frequency of the SVC, SV_{pass} and SV types, the main focus of the subsequent subsections will be on these three syntactic types, while taking into account discipline, NS status and level of achievement. Before moving on to these comparisons, however, two points arising from the study that are more general in character will be addressed.

First, the SVC type is uncontestably the most frequently occurring syntactic type. Nonetheless, the remaining syntactic types still account for approximately one third of the total number of occurrences of the pattern. In particular, the second and third most common syntactic types (SV and SV_{pass}) exhibit relatively high frequencies. This suggests that, while previous studies focusing only on those instances of the introductory *it* pattern that include an adjective phrase have described the use of the bulk of the tokens, a substantial proportion of tokens has remained unanalysed. It would therefore seem useful for future studies to take a more inclusive approach when possible.

The second point concerns differences across regional varieties and potential ramifications for studies including a NS reference corpus. As mentioned in Section 4.1.1, due to the fact that there were no statistically significant differences across the BrE and AmE subcorpora, they were merged into one NS reference corpus along with the NS component of ALEC. The frequencies across the syntactic types for the two varieties were, nonetheless, monitored, and there was one syntactic type for which there was a statistically significant difference between BrE and AmE, namely the SV_{pass} type (27).

(27) *It* must be noted *that speech can still be understood via the left ear too.* (BAWE_LING.6174d)

The difference in frequency of use of the SV_{pass} type between the two regional varieties is striking across both disciplines. Figure 2 shows the normalized frequencies for the SV_{pass} type in the BrE data (checkered bars) and in the AmE data (solid bars); the results for the higher-graded NNS papers (striped bars) are included for reference.

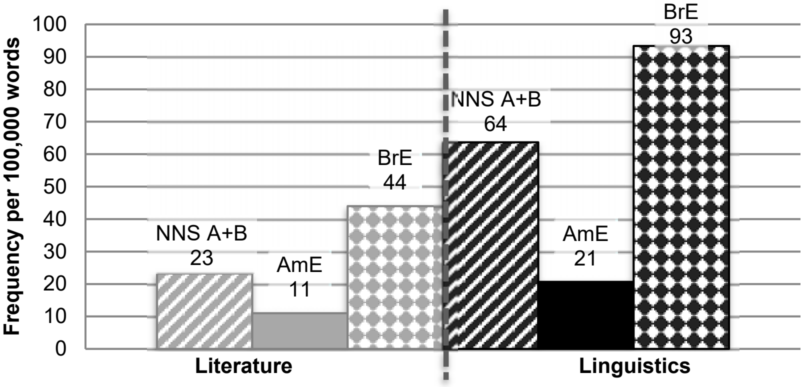


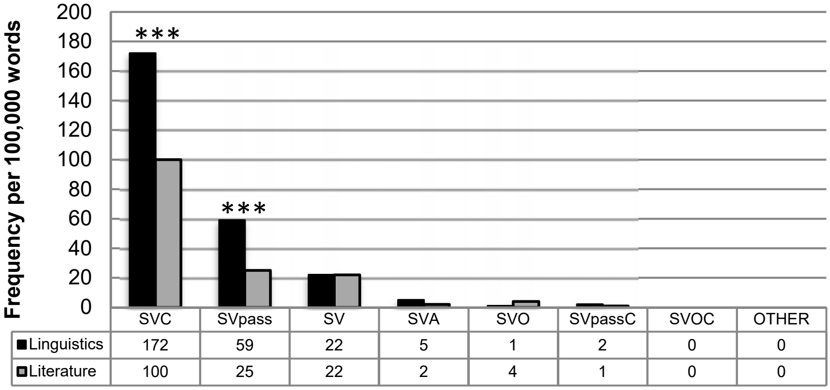
Fig. 2: Frequency (per 100,000 words) of the SV_{pass} type across literature (to the left) and linguistics (to the right).

As shown in Figure 2, while the SV_{pass} type is very frequent in the BrE data, it was comparably infrequent in the AmE data, to the point that it almost seems to be dispreferred. One potential explanation for the difference is offered by Leech et al. (2009: 148–149), who note that there appears to be a stronger prescriptive influence on AmE than on BrE when it comes to preference for the active voice over the passive; however, further investigations of this fall outside the scope of the present study and have to be left for future studies.

Although this finding is based on a relatively small dataset, it brings up the question of which variety (if any) ought to be seen as constituting the standard to which the NNS students should be compared. A choice of one variety over another would result in vastly different results: either the higher-graded NNS students are underusing the SV_{pass} type, or they are overusing it. In this case, however, it actually seems as if the NNS students have been influenced by both BrE and AmE, as the NNS frequencies fall in-between the lower BrE frequencies and the higher AmE frequencies. Still, for investigations using a NS corpus as a reference corpus, this finding suggests that it is preferable for NS reference corpora to include more than one variety of English in order to give a more nuanced picture of NS English.

4.2 Comparison across disciplines

When the frequencies across the two disciplines are compared, it becomes clear that both linguistics and literature exhibit the same overall patterning across the syntactic types as the overall frequencies reported in Section 4.1.2, since the SVC type is the most frequent syntactic type, followed by SV$_{pass}$ and SV. Interestingly, this comparison also shows that the higher overall frequency of occurrence of the introductory *it* pattern in the linguistics texts that was reported in Section 4.1.1 can be explained by the frequencies of the two most frequent syntactic types, SVC and SV$_{pass}$, as shown in Figure 3. The darker-coloured bars show the normalized frequencies of each syntactic type in the linguistics papers and the lighter-coloured bars show the normalized frequencies in the literature papers. The stars represent level of statistical significance across the disciplines.



	SVC	SVpass	SV	SVA	SVO	SVpassC	SVOC	OTHER
■Linguistics	172	59	22	5	1	2	0	0
▨Literature	100	25	22	2	4	1	0	0

Fig. 3: Frequency (per 100,000 words) per syntactic type across linguistics and literature (* = *p*<0.05; ** = *p*<0.01; *** = *p*<0.001).

The fact that there are disciplinary differences is a finding that concurs with both Groom's (2005) and Peacock's (2011) studies on expert writing, as well as with Thompson's (2009) study, in which the overall frequency of the introductory *it* pattern was investigated in NS student writing. However, since the results of the present study show that the linguistics students' more frequent use of the pattern extends to both the overall frequencies and across the two most frequent syntactic types, the question arises what the linguistics students use these syntactic types for that the literature students do not. As the statistically significant differences lie in the SV$_{pass}$ and SVC types, the use of these syntactic types will now be discussed in more detail, followed by a brief discussion of one disciplinary difference found in the SV category that is more qualitative in nature.

As regards the SV_{pass} type, a closer look at the data shows that at least some of the divergence between the linguistics and literature students can be explained by two subtypes, namely *it* + *was* + VERB + *that*-clause and *it* + *has been* + VERB + *that*-clause, as exemplified in (28) and (29), both of which were more frequently used by the linguistics students.

(28) *It* was shown *that translators used a variety of lexemes* [...].
(ALEC_LING.130)

(29) [...] *it* has been found *that skilled readers make less use of context*.
(BAWE_LING.6174e)

In fact, while there were no instances of *it* + *was* + VERB + *that*-clause found in the literature data, 56 occurrences were found in the linguistics data. Moreover, the difference between the two disciplines with regard to *it* + *has been* + VERB + *that*-clause was highly statistically significant.[2] These two subtypes accounted for 38 percent (87/230) of the total number of linguistics tokens categorized as SV_{pass}. Furthermore, there were two verbs whose high frequency in the linguistics texts could in part be explained by their tendency to occur in the verb slot of these subtypes, namely the two verbs found in the examples above: SHOW and FIND. There were no instances of FIND in the literature data (compared to 22 instances in the linguistics data), and SHOW proved to be significantly[3] more frequent in the linguistics data. One explanation for the frequent use of these two subtypes and these verbs involves the fact that the object of study typically differs between linguistics and literature. Unlike most literature papers, the linguistics papers report on empirical studies, in which the students (or previous studies) FIND something and which include results that SHOW something. These results are then reported on using, among others, the two subtypes *it* + *was* + VERB + *that*-clause and *it* + *has been* + VERB + *that*-clause.

For the SVC type, the most common subtype includes those occurrences in which the complement is realized by an adjective phrase rather than a noun phrase (94 percent, 1052/1123, include a complement that is realized by an adjective phrase). Investigating the adjective used for this syntactic type therefore provides a good starting point for further analysis. A closer look at these tokens shows, however, that apart from the linguistics students' slightly more frequent use of each adjective, the two groups used this subtype of SVC in a very similar

[2] $p<0.001$ Pearson's Chi-squared test.

[3] $p<0.01$ Pearson's Chi-squared test.

way, as there is considerable overlap. Table 5 lists the five most common adjectives per discipline.

Tab. 5: The most commonly occurring adjectives across the disciplines.

Rank	Linguistics		Literature	
	Token	Freq. per 100,000 words	Token	Freq. per 100,000 words
1	possible	26	possible	10
2	important	20	clear	9
3	difficult	13	important	8
4	clear	11	impossible	4
5	interesting	10	interesting	3

Unlike the SV$_{pass}$ type, where two subtypes accounted for much of the difference across the disciplines, the difference reported for the SVC type seems to be due to a tendency on the part of the linguistics students to use this syntactic type more frequently overall. As indicated above, one possible explanation for this has to do with potential differences across 'disciplinary cultures' (Hyland 2004: 8–12). The introductory *it* pattern is commonly used to depersonalize claims and to add "a flavour of objectivity and authority to the utterance" (Kaltenböck 2005: 137). Since the linguistics texts involve reports on empirical studies, these texts are likely to put more focus on objective description of the results than the literature texts, as the latter require the author's textual voice to be more clearly detectable. More frequent overall use of the introductory *it* pattern would then be expected in the linguistics texts, as was the case in the present study.

In addition to these differences, there was a difference of more qualitative character in the category that was used equally frequently in both disciplines, the SV type. Here, a use of the pattern that was specific to the literary texts was found: this syntactic type was frequently used by the literature students to comment on claims made about a named character in a work of fiction, as in (30) and (31).

(30) *It* seems *that* *Orsino* *is more concerned with merely being in love* [...]. (MICUSP_LIT.035.1)

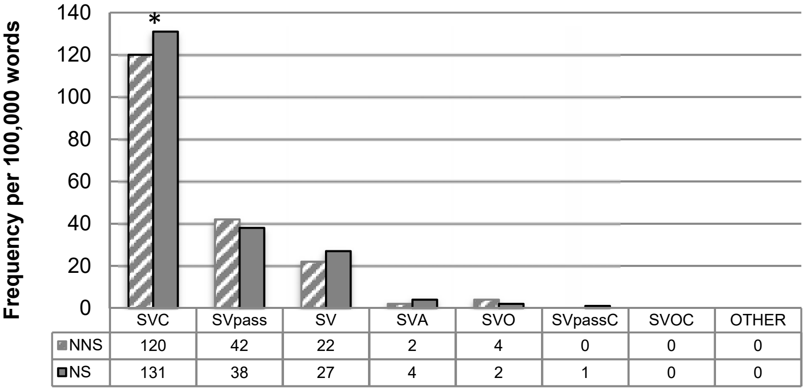
(31) *It* seems *that* *Burton* *suggests not courtly music and refined arts, but rather the popular music of the alehouse* [...]. (ALEC_LIT.065)

The findings suggest that the introductory *it* pattern can be used for purposes that can be considered central to each of the disciplines – be it to depersonalize claims and report on empirical findings in linguistics studies or to comment on named characters in literary works. The pattern could thus serve as a useful tool when teaching students about discipline-specific conventions.

As the disciplinary differences proved to be substantial, the frequencies for the syntactic types across NS status and level of achievement in the following two subsections will be presented both with and without disciplinary differences taken into account, in order to investigate the use of the introductory *it* pattern more extensively.

4.3 Comparison across NS status

Since the NS reference corpus (consisting of subsets of ALEC, BAWE and MICUSP) comprises exclusively higher-graded papers, this part of the study presents the results of a comparison between those papers and the higher-graded NNS papers. As is shown in Figure 4, the frequencies were found to pattern very similarly across the syntactic types.



	SVC	SVpass	SV	SVA	SVO	SVpassC	SVOC	OTHER
▨NNS	120	42	22	2	4	0	0	0
■NS	131	38	27	4	2	1	0	0

Fig. 4: Frequency (per 100,000 words) per syntactic type in the NNS corpus (striped bars) and the NS reference corpus (solid bars).

Only one syntactic type was significantly underused overall (at the 0.05 level) by the high-achieving NNS students: the SVC type. In order to explore the use of this type more thoroughly, the results of an investigation of the use of this syntactic type *per text* are presented below. While the total frequencies provide a valuable overview, the per-text frequencies allow for further (and complementary) investigation of the dispersion. Once an overview has been presented, this kind of investigation allows for potential differences and similarities to be detected also at the level of the individual, which is of importance especially since the per-text results do not necessarily concur with the results of the total frequencies, as we shall see below.

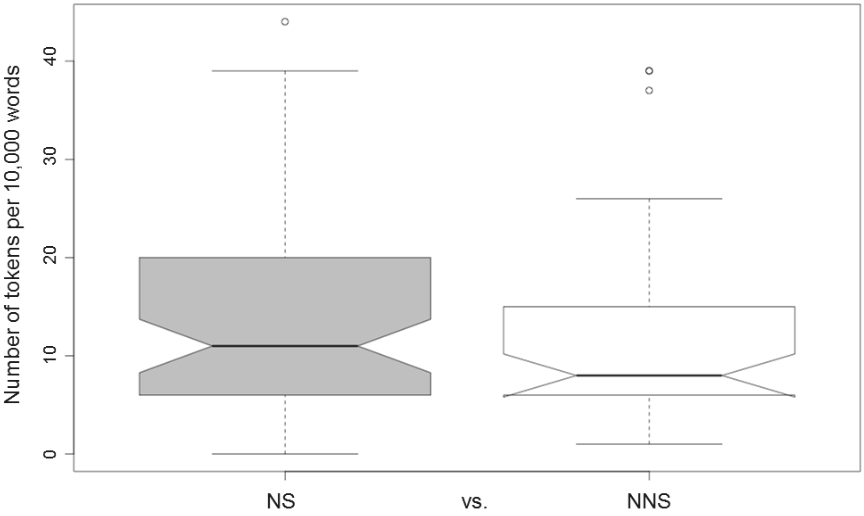
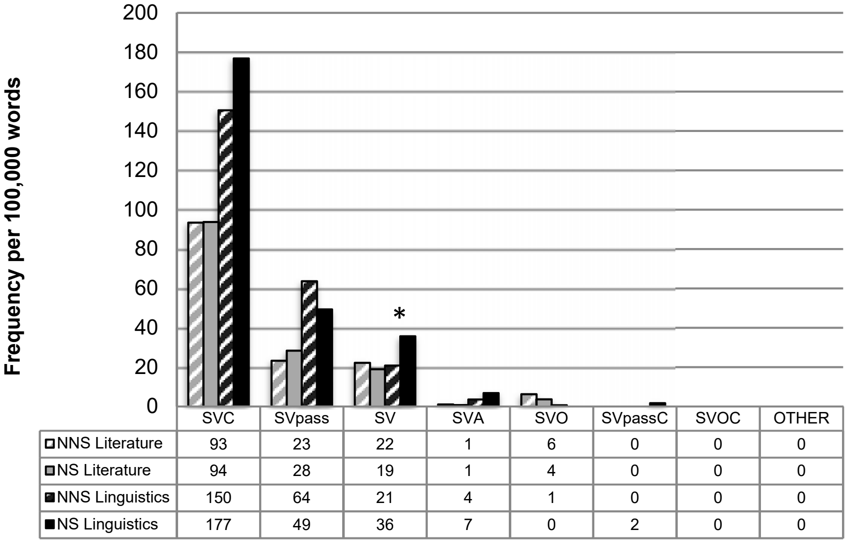


Fig. 5: Frequency (per 10,000 words) of the SVC type per text in the NS (grey) and NNS (white) texts.

In fact, while there was a statistically significant difference at the group level with regard to the use of the SVC type, as reported above, the results for the per-text frequencies show that there were no statistically significant differences between the NS and the high-achieving NNS students' use of the SVC type at the level of the individual, as shown in Figure 5. While bar plots are used in this article to give an overview of the total frequencies across different populations, notched box plots will be used to illustrate the frequencies per text, as this type of graph provides useful information about the dispersion of the frequencies. The results for the NS papers are shown in grey and the results for the higher-graded NNS papers in white. The black lines in the middle of the boxes show the median for the data, and the notches illustrate the confidence interval around the median. If the notches overlap (as is the case here), it indicates that there is no statistically significant difference between the medians; whether this is the case can be tested for significance using, for example, the Kruskal-Wallis rank sum test. The box displays the interquartile range (the 25 percent quartile to the 75 percent quartile). The whiskers (the vertical lines extending downward/upward from the box) show either the minimum/maximum or 1.5 times the interquartile range below/above the 25/75 percent quartile, whichever is larger/smaller. The dots represent outliers.

As is indicated by the overlapping notches in the graph (and as confirmed by a Kruskal-Wallis rank sum test), the higher-graded NNS papers, while exhibiting a slightly smaller range, did not contain significantly fewer SVC tokens than did the NS papers.[4] With these results taken into consideration, it can thus be concluded that the statistically significant difference between the high-achieving NNS students' use of the pattern and the NS students' use was only found at the group level. It is worth noting that this is the case despite the differences between the NNS corpus and the NS reference corpus with regard to text length, background knowledge of the students and the amount of teacher feedback provided to the student, as was noted in Section 3.1.

[4] p>0.05, Kruskal-Wallis rank sum test.



	SVC	SVpass	SV	SVA	SVO	SVpassC	SVOC	OTHER
□NNS Literature	93	23	22	1	6	0	0	0
▨NS Literature	94	28	19	1	4	0	0	0
▩NNS Linguistics	150	64	21	4	1	0	0	0
■NS Linguistics	177	49	36	7	0	2	0	0

Fig. 6: Frequency (per 100,000 words) per syntactic type in the NNS corpus (striped bars) and the NS reference corpus (solid bars) divided up by discipline.

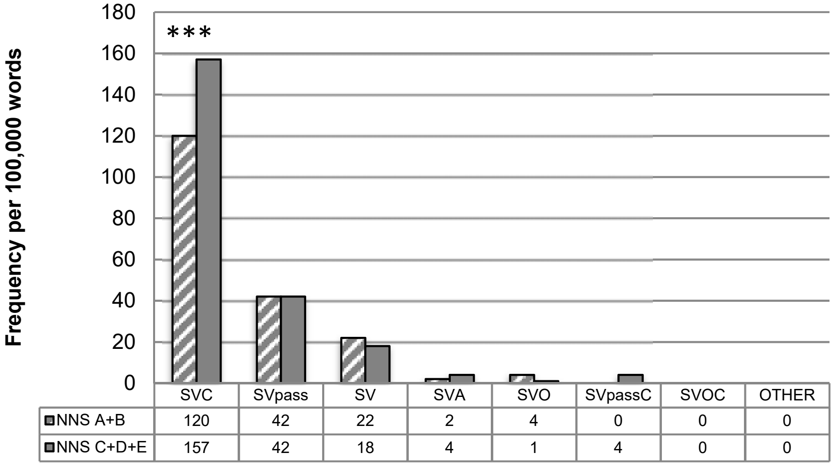
So far in this subsection, the results have been presented without possible cross-disciplinary differences taken into account. In order to investigate whether these similarities in fact characterize *both* linguistics *and* literature, the comparison will now be extended to include separate counts for the two disciplines. Figure 6 shows the normalized frequencies for literature (lighter-coloured bars) and linguistics (darker-coloured bars) respectively.

These results show that even with the disciplines separated, the high-achieving NNS students use the pattern in a way that is not significantly different from the NS students' way of using it, with one exception. The literature texts in particular exhibit remarkably similar normalized frequencies. There is no longer a statistically significant difference across NS status for the SVC type. The only statistically significant difference when the disciplines are separated is the high-achieving NNS linguistics students' tendency to underuse the SV type; however, once again, further analysis showed that the significant difference at group level did not extend to a statistically significant difference at the level of the individual.

All in all, two main points thus emerge from this subsection. First, the high-achieving NNS students seem to be using the syntactic types of the introductory *it* pattern largely in a native-like manner, as witnessed by the similar frequency distributions. Moreover, very few of the differences found proved to be statistically significant. Second, the fact that the differences between the groups that were found to be statistically significant were only found at the group level (and thus not at the level of the individual) emphasizes the importance of also taking individual variability into account. The results thus seem to support claims made by researchers such as Durrant and Schmitt (2009: 168) who claim that the traditional approach of comparing "corpora as wholes" may run the risk of generating "misleading results".

4.4 Comparison across level of achievement

In order to investigate whether the syntactic types are used differently across level of achievement (and thus how this variable may affect the use of the pattern), the higher-graded NNS papers investigated in the previous subsection are here compared to the lower-graded NNS papers (approximately 590,000 words in total). The results of the comparison are presented in Figure 7.



	SVC	SVpass	SV	SVA	SVO	SVpassC	SVOC	OTHER
▨ NNS A+B	120	42	22	2	4	0	0	0
■ NNS C+D+E	157	42	18	4	1	4	0	0

Fig. 7: Frequency (per 100,000 words) per syntactic type in the higher-graded papers (striped bars) and the lower-graded papers (solid bars).

As shown in Table 3 in Section 4.1.1, the lower-graded papers contain significantly more instances of the introductory *it* pattern overall (226 instances per

100,000 words, compared to the higher-graded papers, which contain 189 instances per 100,000 words). From Figure 7, it becomes clear that this difference can be explained by the statistically significant difference in frequency of use of the SVC type. It seems, then, that the two NNS groups exhibit opposite problems: while the students whose papers were awarded a higher grade *undershot* the NS goal for the SVC type slightly at the group level (see Section 4.3), the students whose papers received a lower grade *overshot* the goal considerably at the group level, in comparison both to the NS group and to the high-achieving NNS group. Interestingly, however, in contradistinction to the comparison across NS status, this difference between the two NNS groups remains when the frequencies per essay are explored, as shown in Figure 8.

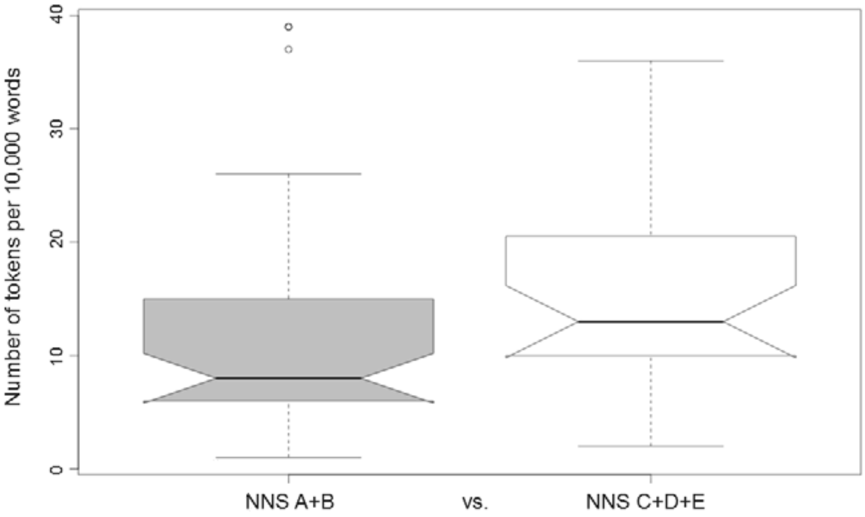


Fig. 8: Frequency (per 10,000 words) of the SVC type per text across the higher-graded papers (grey) and the lower-graded papers (white).

The higher-graded NNS papers contained significantly[5] more instances of the SVC type per text, which, taken together with the group-level frequencies reported in Figure 7, suggests that the students who received a lower grade tend to have a stronger preference for this particular syntactic type. One explanation as to why the lower-graded papers contain such frequent use of the SVC type could have to

[5] $p<0.05$, Kruskal-Wallis rank sum test.

do with it being the most *salient* syntactic type in the academic texts that these students read. The SVC type is by far the most frequently occurring syntactic type across all the corpora included in the analysis and most previous studies have focused exclusively on tokens from this category. It thus seems reasonable to hypothesize that the SVC type is the most frequent syntactic type in academic writing in general. As research on language acquisition has found that the most frequent use of a category is most easily learned (cf., e.g., Ellis, O'Donnell, & Römer 2013) and as the students who received a lower grade can be expected to be slightly less skilled users of English,[6] it could be the case that they are more likely to make "safe choices". These students could therefore be expected to make frequent use of the most prototypical syntactic type (i.e. the type that is likely to be best known to them). If this is indeed the case, then these learners might exhibit what was described by Hasselgren (1994: 256) as a preference for well-known items, which she called 'lexical teddy bears', although '*lexico-grammatical* teddy bears' seems like a more appropriate term in this context.

In order to investigate this hypothesis further, the relative frequencies of the syntactic types as well as the most commonly occurring adjectives inside the SVC tokens were investigated. The results appear to offer some support for the hypothesis. The comparison of the relative frequencies shows that the students whose papers received a lower grade exhibit an especially strong preference for the SVC type in relation to the other syntactic types; the difference was statistically significant.[7] With regard to the most commonly occurring adjectives, both NNS groups frequently used *possible* (32), *important* (33) and *clear* (34). This was especially the case for the lower-graded papers.

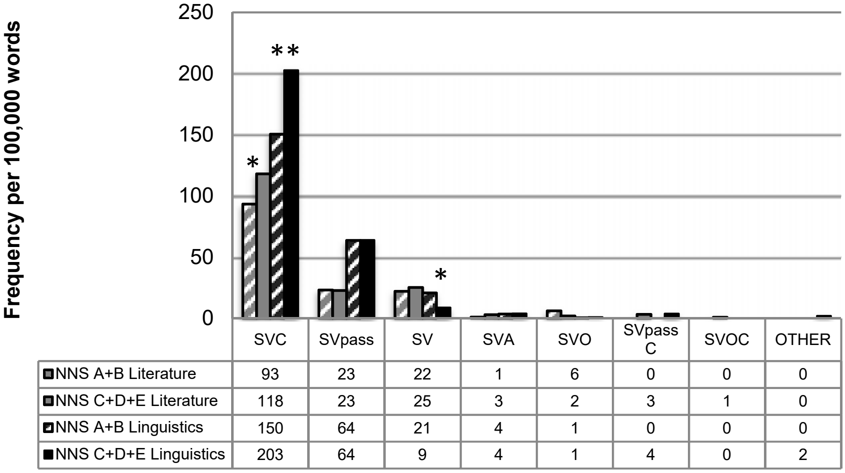
(32) [...] *it* is possible *to draw the conclusion that one can define a pun* [...]. (ALEC_ LING.024)
(33) At this point *it* is important *to clarify what Ndebele means* [...]. (ALEC_ LIT.080)
(34) *It* is clear *that the city has its poisonous grip on him* [...]. (ALEC_LIT.067)

Finally, when the results are broken up into disciplines, it becomes clear that the preference for the SVC type in the lower-graded papers extends to both linguistics and literature. The results of these comparisons are displayed in Figure 9. While

[6] Language-related issues are included in many of the criteria on which the students' papers have been assessed.

[7] $p<0.05$, Pearson's Chi-squared test.

there also appears to be a difference between the linguistics subcorpora with regard to the SV type, a closer look at the dispersion shows that this difference is, in fact, due to an outlier.



	SVC	SVpass	SV	SVA	SVO	SVpass C	SVOC	OTHER
▨NNS A+B Literature	93	23	22	1	6	0	0	0
■NNS C+D+E Literature	118	23	25	3	2	3	1	0
▨NNS A+B Linguistics	150	64	21	4	1	0	0	0
■NNS C+D+E Linguistics	203	64	9	4	1	4	0	2

Fig. 9: Frequency (per 100,000 words) per syntactic type in the higher-graded papers (striped bars) and the lower-graded papers (solid bars), divided up by discipline.

Overall, unlike the slight differences across NS status reported in Section 4.3, the substantial differences in the use of the SVC type across level of achievement presented in this section proved to extend not only to the group level, but also to the level of the individual and across both disciplines. When the results were examined in more detail, the differences with regard to the use of the SVC type were also found to be statistically significant for the relative frequencies across level of achievement. It thus seems that the lower-graded students in particular would benefit from explicit teaching of alternative ways to use the pattern, along with other patterns, in order for them not to cling too much to linguistic teddy bears of any sort. These findings furthermore highlight the fact that it is not necessarily the case that frequent use of the introductory *it* pattern can be equated to proficient use of it.

5 Conclusion

Starting out from Quirk et al.'s (1985: 1392) seven syntactic types of the introductory *it* pattern, this study has not only investigated what constitutes the full inventory of these syntactic types in academic writing by university students, but has also mapped out the frequency of occurrence of each of the types. Comparisons were made across academic disciplines, NS status and level of achievement. Overall, the results showed that the frequencies of the syntactic types pattern similarly, as the relative order for the three most frequently occurring syntactic types was the same across all the subcorpora. This suggests that the use of the pattern is very stable in both NS and NNS student writing. It furthermore became clear that the introductory *it* pattern is not a monolithic pattern. While the SVC type (especially the subtype including an adjective phrase) is the most frequently occurring syntactic type in all subcorpora, the SV and SV$_{\text{pass}}$ types also exhibit relatively high frequencies. These results thus highlight the importance of not limiting the discussion of the introductory *it* pattern to only tokens which include an adjective phrase, as has been the approach of many previous studies, since such analyses would not give the full picture.

Interestingly, the results for all subcorpora display clear differences between linguistics and literature; the use of the introductory *it* pattern thus seems to be highly discipline-specific. For example, while the literature students commonly use the SV type of the pattern to evaluate claims involving a character in a work of fiction, the linguistics students make significantly more frequent use of certain subtypes of the SV$_{\text{pass}}$ type, such as *it* + *has* + *been* + VERB + *that*-clause to report on empirical findings. The introductory *it* pattern, then, seems to be used to perform tasks that are central to these disciplines, which suggests that the pattern deserves a place among discipline-specific conventions that should be taught to students.

Although some frequency differences were noted when the higher-graded NNS papers were compared to the NS student papers, the NNS students were found to use the pattern in a native-like manner overall, which is in line with previous research (Römer 2009). More noteworthy differences were instead found when the two NNS corpora were compared. Here, the use of the syntactic types was compared across level of achievement, a factor that has, to the best of my knowledge, not previously been investigated in relation to the use of the introductory *it* pattern. This comparison showed that the NNS students whose paper received a lower grade exhibited a particularly strong preference for the most frequent syntactic type, the SVC type, in a way that suggests that these students tend to cling to what could be referred to as a 'lexico-grammatical teddy bear' (cf. Hasselgren's 1994 use of 'lexical teddy bears'). It would then seem beneficial for these students in particular to be made aware of other variants of the introductory *it* pattern as well. These findings furthermore show that one cannot necessarily equate frequent use of the pattern with proficient use of the pattern, a topic on which more research is needed.

Further possible avenues for future research include a large-scale investigation of the use of the introductory *it* pattern in learner writing and expert writing. One of the advantages of using NS student writing as a reference is that this presents the learners with a more attainable goal; just as the NS students cannot be expected to be expert writers, it does not seem fair to expect the learners to live up to expert standards either. Nevertheless, additional comparisons with expert writing would most certainly provide further insights into the use of the introductory *it* pattern in academic discourse.

Finally, by highlighting the versatility of the introductory *it* pattern in terms of both syntactic make-up and function, teachers could make their students aware of the usefulness of the pattern in academic writing. Since discipline proved to be an important predictor for the use of the introductory *it* pattern, increasing students' awareness of the pattern would be one way of helping students develop their academic "voice", which would enable them to communicate successfully with the research community to which they aspire to belong.

References

Advanced Learner English Corpus (ALEC). (2013). Corpus compiled by Tove Larsson at Uppsala University, Sweden.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

British Academic Written English Corpus (BAWE). (2004—2007). Corpus compiled at the Universities of Warwick, Reading and Oxford Brookes, UK. Retrieved from http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/ (accessed April 2016, last accessed October 2018).

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/ (accessed April 2016, last accessed October 2018).

Davies, M. (2008–). *The Corpus of Contemporary American English (COCA): 520 million words, 1990–present*. Retrieved from: http://corpus.byu.edu/coca/ (accessed April 2016, last accessed October 2018).

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition. *Currents in Language Learning, 63*(1), 25–51.

Francis, G., Manning, E., & Hunston, S. (1998). *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London, UK: HarperCollins.

Granger, S. (2002). A birds-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching* (pp. 3–33). Amsterdam, Netherlands: John Benjamins.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, *4*(3), 257–277.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–260.

Heuboeck, A., Holmes, J., & Nesi, H. (2008). *The BAWE corpus manual*. Retrieved from http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf (accessed March 2014, last accessed October 2018).

Hewings, M., & Hewings, A. (2002). "It is interesting to note that…": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, *21*(4), 367–383.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor, MI: University of Michigan Press.

Jenkins, J. (2006). Current perspectives on teaching World Englishes and English as a Lingua Franca. *TESOL Quarterly*, *40*(1), 157–181.

Jenkins, J. (2009). English as a lingua franca: Interpretations and attitudes. *World Englishes*, *28*(2), 200–207.

Kaltenböck, G. (2005). *It*-extraposition in English: A functional view. *International Journal of Corpus Linguistics*, *10*(2), 119–159.

Larsson, T. (2016). The Introductory *It* Pattern in Academic Writing by Non-Native-Speaker Students, Native-Speaker Students and Published Writers: A Corpus-Based Study. Doctoral dissertation. Department of English, Uppsala University, Sweden.

Larsson, T. (2017). A functional classification of the introductory *it* pattern: Investigating academic writing by non-native-speaker and native-speaker students. *English for Specific Purposes*, 48, 57–70.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge, UK: Cambridge University Press.

Mair, C. (1990). *Infinitival Complement Clauses: A Study of Syntax in Discourse*. Cambridge, UK: Cambridge University Press.

Michigan Corpus of Upper-level Student Papers. (2009). Ann Arbor, MI: The Regents of the University of Michigan. Retrieved from from http://micusp.elicorpora.info (accessed April 2016, last accessed October 2018).

Mindt, I. (2011). *Adjective Complementation: An Empirical Analysis of Adjectives Followed by That-clauses*. Amsterdam, Netherlands: John Benjamins.

Mukherjee, J. (2006). Corpus linguistics and English reference grammars. In A. Kehoe & A. Renouf (Eds.), *The Changing Face of Corpus Linguistics* (pp. 337–354). Amsterdam, Netherlands: Rodopi.

Peacock, M. (2011). A comparative study of introductory *it* in research articles across eight disciplines. *International Journal of Corpus Linguistics*, *16*(1), 72–100.

Powers, D. A., & Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*. San Diego, CA: Academic Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London, UK: Longman.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/ (accessed April 2016, last accessed October 2018).

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*(1), 140–162.

Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*(2), 159–177.

Scott, M. (2012). WordSmith Tools version 6. Liverpool, UK: Lexical Analysis Software.

Thompson, P. (2009). Shared disciplinary norms and individual traits in the writing of British undergraduates. In M. Gotti (Ed.), *Commonality and Individuality in Academic Discourse* (pp. 53–82). Bern, Switzerland: Peter Lang.

Zipf, G. K. (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.

Hilde Hasselgård

Phraseological teddy bears

Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English

Abstract: This chapter compares frequent four-word lexical bundles in a learner corpus (VESPA) and a native speaker corpus (BAWE), both representing novice academic writing. The frequencies and dispersion of bundles in the two corpora reveal patterns of both over- and underuse among the learners. The learners are shown to use some bundles very frequently, but frequencies drop more sharply than in the native corpus. The dispersion of the frequent bundles tends to be broader in the native speaker corpus. In a closer scrutiny of four selected bundles the novice-expert dimension is addressed by consulting a corpus of published research articles. Contrasts between English and Norwegian are also considered in order to explain the learners' apparently non-native usage. Some of the most overused bundles seem to have been generalized by the learners to fit into contexts where native speakers rarely use them; these can be described as 'phraseological teddy bears'. Pedagogical applications of the results should start from the underused items in order to broaden the phraseological repertoire of the learners.

1 Introduction

It is well established that learners as well as native speakers use pre-fabricated multi-word units in their language production (see e.g. Granger 1998). Yet, "phraseology is one of the aspects that unmistakably distinguishes native speakers of a language from L2 learners" (Granger & Bestgen 2014: 229), and the phraseology of non-native users of English continues to inspire investigations into the puzzle of nativelike co-selection (Pawley & Syder 1983).

The present investigation concerns the most frequent four-word lexical bundles in two corpora of novice academic English representing advanced learners (with Norwegian as their L1) and native speakers of English. The bundles most frequently used by the two writer groups will be compared with regard to their

Hilde Hasselgård, University of Oslo, hilde.hasselgard@ilos.uio.no

https://doi.org/10.1515/9783110489071-013

distribution, meanings and functions. I will also take a closer look at some selected bundles whose frequencies and distributions differ markedly between the corpora.

Ringbom (1998) shows that the frequencies of individual word forms tend to differ between learners and native speakers of English, with learners having a tendency to overuse vocabulary items that have high frequencies in general corpora of English. The overuse can be related to a core vocabulary that the learners have acquired early and know well. Hasselgren (1994: 237) compares such familiar lexical favourites to children's toys: "[s]tripped of the confidence and ease we take for granted in our first language flow, we regularly clutch for the words we feel safe with: our 'lexical teddy bears'". A hypothesis of the present study is that the same tendency will be visible in the use of lexical bundles: some bundles will seem familiar and unobjectionable to learners, who will resort to them frequently as their '*phraseological* teddy bears'. This idea is not novel; Nesselhauf (2005: 247) suggests that learners' occasional overuse of "certain native-speaker-like chunks" may partly result "from learners using some of them as lexical teddy bears".[1] Other bundles, however, will be underused by learners, for example because most learners simply do not know them, or because they belong to a style level that the learners are not fully familiar with. At the advanced level of proficiency represented in the learner corpus used (see Section 3 for details), the differences between native and non-native usage of bundles are not expected to consist in errors as much as in diverging frequencies of use.

Hasselgren (1994: 237–238) seems to imply that words characterized as lexical teddy bears are not only more frequent in learner language than in native language; they are also "systemically overgeneralized by advanced learners", which leads to their being used in contexts where native speakers would choose a (near) synonym (see also Levenston & Blum 1977). Thus, a phraseological teddy bear will be a multi-word unit that learners use more frequently and in more contexts than native speakers do.

The chapter is structured as follows: after a review of relevant previous research and a presentation of material and method, the most frequent lexical bun-

[1] Ellis (2012: 29) uses the term 'phrasal teddy bear' to refer to "highly frequent and prototypically functional phrases like *put it on the table, how are you?, it's lunch time*", or "formulaic phrases with routine functional purposes" (2012: 37). Since lexical bundles, unlike Ellis's formulaic sequences, do not require word strings to be idiomatic or complete functional units (Biber et al. 1999: 990), I have opted for the related term 'phraseological teddy bear'.

dles in both corpora will be identified and discussed. Then follow four case studies of selected bundles that are either overused or underused by the learners before some concluding remarks are offered.

2 Some previous studies of lexical bundles and formulaic language in learner English

Recurrent strings of words have been studied under a number of different headings, for example ‘recurrent word combinations’ (e.g. Altenberg 1998), ‘n-grams’ (e.g. Granger & Bestgen 2014; Ebeling & Hasselgård 2015a), and ‘lexical bundles’ (e.g. Biber et al. 1999; Cortes 2004; Ädel & Erman 2012; Paquot 2013). For general overviews of phraseology in learner corpus research, see Paquot and Granger (2012) and Ebeling and Hasselgård (2015b).

Lexical bundles are defined as “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (Biber et al. 1999: 990). The operationalization of the definition limits lexical bundles to “uninterrupted combinations of words” that occur above a set frequency threshold and across a minimum number of corpus texts “to exclude individual speaker/writer idiosyncrasies” (1999: 993). Biber et al. show that conversation and academic prose differ in their use of bundles as regards lexicogrammatical structure as well as frequency (e.g. 1999: 997).

Using a similar method, but the term ‘chains’, Stubbs and Barth (2003) show that recurrent phrases can be used as text type discriminators. That is, they identify differences between the text types by applying a number of measures, one of which is “recurrent word-chains and/or their comparative frequency” (2003: 79). Cortes (2004) discusses lexical bundles in the academic disciplines history and biology. She classifies the lexical bundles functionally into ‘referential bundles’, ‘text organizers’ and ‘stance bundles’ (2004: 409), and shows that the academic disciplines history and biology vary in their use of lexical bundles, in terms of both structural and functional features. She also finds “that the use of target bundles by students in biology and history courses at different university levels very far from resembles the use of these bundles by published authors in these disciplines” (2004: 421). Hyland (2008) similarly shows that there is variation between published academic writing and postgraduate student writing, and furthermore, that bundle usage differs across academic disciplines.

The functional classification in Cortes (2004) is also used by Biber, Conrad and Cortes (2004), in a paper much referred to in subsequent research on lexical bundles in academic language. It shows clear differences as regards structural and functional categories of lexical bundles across four ‘university registers’: conversation, classroom teaching, textbooks and academic prose. Importantly the authors conclude that “lexical bundles should be regarded as a basic linguistic construct with important functions for the construction of discourse” (2004: 398).

The study of lexical bundles is extended to a comparison of learners and native speakers in Chen and Baker (2010), who compare the writing of Chinese learners of English to native speaker student and expert writing.[2] They find “that the use of lexical bundles in non-native and native student essays is surprisingly similar” while professional writing shows a wider repertoire of certain types of bundles (2010: 44). However, this applies mainly to the quantitative analysis; the qualitative analysis reveals some differences between native and non-native writing. An interesting observation for the present study is that non-native writing tends to show features of “over-generalizing and favoring certain idiomatic expressions and connectors” (Chen & Baker 2010: 44). Ädel and Erman (2012), in a study that to some extent replicates Chen and Baker (2010) with data from Swedish learners of English, find more substantial differences between native and non-native writing. They conclude that “non-native speakers exhibit a more restricted repertoire of recurrent word combinations than native speakers” (Ädel & Erman 2012: 90). The qualitative analysis of context is singled out as a future direction in the study of lexical bundles in learner language, since the fact that a bundle may be used to the same extent by learners as by native speakers does not necessarily entail “that it is used in the same way” by both groups (Ädel & Erman 2012: 91).

As for the use of lexical bundles in Norwegian-produced learner English, the present study has a precursor in Lie (2013), who examines the use and functions of bundles containing three or more words. He finds that Norwegian learners use lexical bundles for much the same functions as native speakers, but rely on a smaller repertoire of bundles, sometimes overgeneralizing their meaning and use (Lie 2013: 47). There is also a tendency among learners to prefer less formal alternatives to more academic ones (Lie 2013: 48).

Pérez-Llantada (2014) focuses on 4-word bundles in L1 and L2 expert academic writing. Importantly she correlates her findings of L2 English with bundles

[2] Chen and Baker (2010) used subsets of the British Academic Written English corpus (BAWE) for the learner and native speaker student comparison. BAWE comprises English texts from a number of L1 backgrounds besides English, with Chinese being the most frequent non-English L1; see Nesi and Gardner (2012: 268).

in the writers' first language, Spanish. A central conclusion is that "the L2 English variable reflects a 'hybrid' formulaic language" (Pérez-Llantada 2014: 92); i.e. it is not fully nativelike and shows traces of transfer from L1 Spanish phraseology. The use of bundles in L1 and L2 expert academic writing is also the topic of Salazar (2014) who studies the frequency, structure and function of 3–6-word bundles extracted from corpora of biomedical research writing. As Salazar's (2014: 46, 153) purpose is partly to explore pedagogical applications, bundles are selected according to a combination of frequency and Mutual Information (MI) to produce a list of pedagogically relevant bundles.

Ebeling and Hasselgård (2015a) compare the use of n-grams across academic disciplines and L1 backgrounds in VESPA and BAWE, concluding that both factors have an impact on n-gram use, although discipline seems to be the stronger cause of differences. This study concerns functional types of n-grams, classified in line with Moon (1998) as ideational, interpersonal and textual (a framework similar to the one found in Cortes 2004 and Biber, Conrad, & Cortes 2004). The study does not focus on the frequencies of individual n-grams in the corpora, but as in Ädel and Erman (2012) one of the envisaged avenues of further research is a more qualitatively oriented study which takes token frequency and context into account. The present study can be seen as a step in that direction and an attempt to fill a gap in present research.

3 Material and method

Two corpora form the core material for the present investigation: the Norwegian component of the **V**arieties of **E**nglish for **S**pecific **P**urposes d**A**tabase (VESPA-NO) and the **B**ritish **A**cademic **W**ritten **E**nglish corpus (BAWE). Both corpora contain student writing within a variety of academic disciplines. For the present purposes only the linguistics discipline has been investigated, and only texts written by students whose L1 is Norwegian and English, respectively. Table 1 shows the size and composition of the corpora used.[3]

[3] This study is based on the 2012 version of VESPA-NO. The corpus has been updated and slightly enlarged since then.

Tab. 1: The two main corpora for the study.

Discipline = linguistics	Texts	Words
VESPA-NO (L2)	239	267,855
BAWE (L1, BrE)	76	167,437

Both corpora have been annotated in order for searches to ignore material not produced by the student, such as linguistic examples, quotations and bibliographies. The word counts given in Table 1 are exclusive of the ignored material. See Ebeling and Heuboeck (2007) and the respective corpus manuals (Paquot et al. 2010; Heuboeck, Holmes, & Nesi 2008) for more information regarding the annotation.

In the discussion of bundles selected for the case studies in Section 5, I also draw on other corpora. The Corpus of Research Articles (CRA) held at Hong Kong Polytechnic University will be used to check whether there are differences between novice and expert writers within the same discipline.[4] To investigate potential influence from Norwegian, I will consult the English–Norwegian Parallel Corpus (ENPC) and the KIAP corpus (Cultural Identity in Academic Prose), which contains published research articles in English, Norwegian and French. From KIAP I use only the section containing linguistics articles in Norwegian, comprising 269,913 words (Fløttum, Dahl, & Kinn 2006: 7; Fløttum et al. 2013). Other corpora used for occasional reference are the Michigan Corpus of Upper-Level Student Papers (MICUSP), the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA).[5]

The study takes a lexical-bundle approach (Biber et al. 1999; Biber, Conrad, & Cortes 2004). Lexical bundles are recurrent uninterrupted sequences of word forms that occur above a certain frequency threshold and with a certain dispersion across texts, operationalized in Biber et al. (1999: 992) as at least ten times per million words and in at least five texts. Although lexical bundles may in principle consist of any number of words (above one), the present study is limited to four-word bundles. This decision is much in line with comparable previous studies (e.g. Hyland 2008; Ädel & Erman 2012; Pérez-Llantada 2014). As Hyland (2008: 8) says, “they are far more common than 5-word strings and offer a clearer range

[4] The corpus contains published research articles in a variety of disciplines. This study uses Applied Linguistics (170,653 words), which was considered closer than Linguistics to the topics contained in VESPA and BAWE.

[5] For further information about the corpora, see the websites listed at the end of this chapter.

of structures and functions than 3-word bundles". Some of the resultant 4-word bundles (see Section 4) may be said to consist of a 3-word bundle plus a common word, for example *the meaning of the*, of which the most "salient" part (Simpson-Vlach & Ellis 2010: 490) is arguably *the meaning of*. However, as argued by Hunston (2008), "small words" play an important role in the identification of grammar patterns, which in turn can form semantic sequences (2008: 271), or in the words of Pérez-Llantada (2014: 86), "[b]undles bridge structural units in the discourse, framing semantic meanings". For example, it may be a salient feature of sequences such as *the meaning of* and *the use of* that they occur in the context of an extended noun phrase.

Recurrent four-word bundles were extracted by means of WordSmith Tools 6 (Scott 2012). The focus is on the highest frequency band, i.e. bundles that account for at least 0.01% of the corpus according to WordSmith's Wordlist tool. All the bundles in this frequency band occurred in at least five different texts (cf. Biber et al. 1999: 993; Biber, Conrad, & Cortes 2004: 376). The issue of overlapping bundles (cf. Simpson-Vlach & Ellis 2010: 493) has not been addressed in the present analysis. The potentially overlapping bundles are found in VESPA (see Table 2) and are *is an example of / an example of this* and *the use of the / is the use of*. The overlap concerns one occurrence of *is an example of this* and eight occurrences of *is the use of the*.

The bundles were classified functionally along the lines of Cortes (2004) and Biber, Conrad and Cortes (2004). Although many of the bundles are structurally incomplete, they contain enough meaning-bearing elements to make such classification possible. The categories outlined in Biber, Conrad and Cortes (2004) are the following:

- Referential bundles (R) "make direct reference to physical or abstract entities, or to the textual context itself" (2004: 384)
- Stance bundles (S) "express attitudes or assessments of certainty" (2004: 384)
- Discourse organizers (D) "reflect relationships between prior and coming discourse" (2004: 384)

Bundles that were considered specific to particular topics or tasks were excluded, as they would be unlikely to occur in other corpora. Examples are *Australian and New-Zealand English, Norwegian learners of English, the second text is, as in Tager and Flusberg*.

4 Corpus analysis

The results of the search for frequent four-word bundles in the two corpora are presented in Table 2 along with a functional label and their frequency per 100,000 words. The bundles that are shared between the corpora are marked in shaded cells. The table has fewer bundles from VESPA than from BAWE, but it may be noted that more task- and topic-specific bundles have been removed from the original VESPA list than from the BAWE list (8 vs. 1; see selection criteria in Section 3).

Tab. 2: Four-word bundles in both corpora (frequencies per 100,000 words).

Bundles in VESPA	Function	Freq.	Bundles in BAWE	Function	Freq.
on the other hand	D	38.1	*it is important to*	S	19.1
the use of the	R	32.9	*in the case of*	D	17.3
when it comes to	D	18.7	*as a result of*	D	15.5
the meaning of the	R	18.3	*the use of the*	R	14.9
the rest of the	R	17.2	*to be able to*	R	14.9
is an example of	R	16.8	*the way in which*	R	13.7
an example of this	R	13.8	*the fact that the*	D	11.9
the fact that the	D	13.4	*the way we speak*	R	11.3
is the use of	R	12.3	*can be found in*	R	10.8
as we can see	D	11.2	*on the other hand*	D	10.8
I have chosen to	S	10.5	*it was found that*	R	10.2
in the case of	D	10.5	*the context of the*	R	10.2
			the meaning of the	R	10.2
			to look at the	R	10.2

Two of the bundles that occur in both corpora, *in the case of* and *on the other hand*, are "the most common four-word lexical bundles in academic prose" according to Biber et al. (1999: 994), and apparently the only ones to reach a frequency above 100 per million words. It may be noted that the BAWE list overlaps slightly more than the VESPA list with that presented by Byrd and Coxhead (2010: 37–39) of widely used lexical bundles in the AWL [Academic Word List] corpus, thus suggesting that BAWE bundles are more academic.

The distribution of functional types of bundles is fairly similar between the corpora: VESPA has six referential bundles, five discourse bundles and one stance bundle, while BAWE has nine referential bundles, four discourse bundles and one stance bundle. VESPA has one discourse bundle and one stance bundle that are personal and self-referential (*I have chosen to* and *as we can see*), while BAWE has one personal referential bundle (*the way we speak*).

It is striking that the highest frequencies in VESPA far exceed those in BAWE; there is also a much steeper decline of frequencies in VESPA. This suggests that learners tend to re-use a small number of bundles to a greater extent than native speakers. To investigate overuse and underuse, the token frequencies of all the bundles included in Table 2 were compared between the corpora. A number of bundles had similar frequencies in VESPA and BAWE (*an example of this, the fact that the, in the case of, can be found in*), while the rest differed significantly in frequency according to a log-likelihood test.[6] Table 3 displays those bundles that are either overused or underused in VESPA compared to BAWE.

Tab. 3: Overused and underused bundles in VESPA (raw frequencies).

	Overuse			**Underuse**		
		VESPA	**BAWE**		**VESPA**	**BAWE**
p≤0.0001	*on the other hand*	102	18	*as a result of*	8	26
	when it comes to	50	0	*the way we speak*	0	19
	the rest of the	46	4	*it was found that*	0	17
	is an example of	44	9			
	is the use of	33	3			
	as we can see	30	0			
p≤0.001	*the use of the*	87	25	*it is important to*	18	32
	I have chosen to	30	3	*the way in which*	9	23
p≤0.01				*to be able to*	16	25
p<0.05	*the meaning of the*	47	17	*the context of the*	7	17

In a next step, the dispersion of the most frequent bundles was studied to check whether the frequency of any bundle is boosted because of popularity in certain texts. Table 4 shows the percentage of texts in which each bundle occurs. Note that the frequency order differs slightly from that in Table 2. The most common

[6] Log Likelihood was calculated with Paul Rayson's calculator available at http://ucrel.lancs.ac.uk/llwizard.html (accessed 3 November 2015, last accessed October 2018).

bundle in VESPA has a much wider dispersion than any of the bundles in BAWE, but there is a sharp drop already at rank 2. In other words, most of the frequent bundles in BAWE are more widely dispersed than those in VESPA.

Tab. 4: The most widely used bundles in both corpora (distribution across texts).

Bundles in VESPA	Texts	Bundles in BAWE	Texts
on the other hand	31.0%	*it is important to*	23.7%
the use of the	19.0%	*the fact that the*	23.7%
the rest of the	14.2%	*to be able to*	22.4%
the fact that the	13.4%	*in the case of*	19.7%
is an example of	12.6%	*on the other hand*	18.4%
the meaning of the	11.7%	*as a result of*	17.1%
as we can see	11.3%	*the meaning of the*	17.1%
when it comes to	10.5%	*the use of the*	15.8%
is the use of	10.0%	*the way in which*	15.8%
an example of this	9.6%	*the context of the*	15.8%
I have chosen to	9.6%	*it was found that*	13.2%
in the case of	7.9%	*to look at the*	13.2%
		can be found in	10.5%
		the way we speak	7.9%

A comparison of frequencies based on dispersion gives a different perspective on overuse and underuse. Relating the number of texts each bundle occurs in to the total number of texts in each corpus, the following bundles had similar dispersions in the two corpora: *the rest of the, is an example of, an example of this, is the use of, I have chosen to, the use of the, the meaning of the, the fact that the, can be found in.* The bundles that occur in significantly different numbers of texts in the corpora (according to a chi square test) are shown in Table 5.

When dispersion is taken into account, the number of overused bundles is greatly reduced (compare Tables 3 and 5), while the underuse is relatively unchanged. That is, some of the 'overused' bundles in Table 3 are overused only in some texts. Interestingly, all the overused bundles in Table 5 are discourse organizers, but most of the underused ones are referential.

Tab. 5: Overused and underused bundles in VESPA according to text dispersion.

	Overuse	**Underuse**
p≤0.0001		*it is important to* *as a result of* *to be able to* *the way in which* *the way we speak* *it was found that* *the context of the*
p≤0.001	-	-
p≤0.01	*when it comes to* *as we can see*	*in the case of*
p<0.05	*on the other hand*	

5 Case studies

This section presents case studies of four discourse-organizing lexical bundles whose distributions differ significantly between the corpora as regards both frequency and dispersion. Three of these are overused in VESPA compared to BAWE (*on the other hand, when it comes to, as we can see*), and one is underused (*as a result of*). I will draw up a usage profile for each bundle on the basis of VESPA concordances, and possible reasons for the attested overuse/underuse will be discussed, such as the presence of a corresponding expression in Norwegian, along the lines of Paquot's (2013) study of transfer effects. Unlike Paquot, I have not investigated learner behaviour in EFL corpora with other L1 backgrounds. However, the novice writers in VESPA and BAWE will be compared with 'expert' writers, represented in the applied linguistics section of the CRA, to control for any novice-expert differences.

5.1 *On the other hand*

The most frequent four-word bundle in VESPA is *on the other hand* (see Table 2); it is also the bundle with the widest dispersion, occurring in 31% of the texts.

VESPA also overuses the bundle in comparison with CRA, where there are 22 occurrences (12.9 per 100,000 words). The VESPA concordance shows that *on the other hand* has the following characteristics:

- 30 out of 102 occurrences are clause-initial (example 1); the remaining 72 are clause-medial (example 2).
- *On the other hand* co-occurs with *on the one hand* eight times in VESPA; see example (3). This pattern is not found in BAWE.[7]
- *On the other hand* sometimes functions as a general topic shifter in VESPA, not always marking contrast (Lie 2013); see example (4).
- No extended phraseological pattern can be identified for the bundle.

(1) *On the other hand*, the overuse of the progressive is intralingual in that it reflects what has been learned and has been overgeneralized. (VESPA)[8]

(2) Coherence, *on the other hand*, is in the mind of the writer and reader: it is a mental phenomenon and cannot be identified or quantified in the same way as cohesion. (VESPA)

(3) *On the one hand*, contrastive linguists are very enthusiastic about what they have to offer L2 teaching; *on the other hand*, they seem somehow depressed by the lack of positive response among teachers. (VESPA)

(4) If it works you have substitution, and if we apply this test for this line, we find that it does. “We are the *ones*..” So far, so good. *On the other hand*, what is important to notice here is that we do not really know what the word is substituting for. (VESPA)

The preference for placing the bundle in clause-medial position is found in BAWE, too,[9] with the same discourse effect, namely to set off the subject (or any other clause-initial element) as contrasting with the preceding context. Example (2), for instance, follows directly after a definition of ‘cohesion’, which makes it appropriate to steer contrastive focus to the related, but different, concept. Interestingly, the preferred position of *on the other hand* in CRA is initial (18 out of the 22 occurrences). Three medially placed instances provide contrastive focus to the clause subject, while the remaining one, example (5), occurs in an elliptical clause and gives extra focus to the final constituent. This usage is not found in BAWE or VESPA.

[7] Byrd and Coxhead (2010: 46) also note that “*on the other hand* is most often used [...] without the prior use of *on the one hand*”.

[8] All corpus examples have been rendered as they occur in the corpora.

[9] In BAWE, 12 out of the 18 instances occur in medial position.

(5) In order to summarize the data, Fig. 9 provides a schematic representation of the main confusions identified from the perceptual results, for the adults and the 8-year old, on the one hand, and *on the other hand*, for the 4-year old. (CRA)

The Norwegian expression corresponding most closely to *on the other hand* with respect to similarity of form is *på den annen side* ('on the other side'). This expression is, however, not nearly as frequent in the Norwegian corpora consulted as the frequency of *on the other hand* in VESPA might suggest: there are 6.8 occurrences per 100,000 words in ENPC non-fiction and 9.6 in KIAP (ling.).[10] The overuse in VESPA thus cannot be explained by means of direct transfer from Norwegian, but the positional preference can: all the occurrences of *på den annen side* in KIAP occur in medial position.

Studying the wider context of *on the other hand* in VESPA, we find that it often occurs in the vicinity of other markers of contrast, as exemplified in (6).

(6) The use of cataphoric reference is not very extensive, *however*. Anaphoric reference *on the other hand*, is utilized throughout the text. (VESPA)

This suggests that the learners may have a tendency to over-express contrastive relations when the discourse moves from one topic to another. As noted above, and illustrated in (4), "VESPA contributors are as likely to use the phrase as a general topic change marker, introducing a concept only tangentially related to the preceding sentence" (Lie 2013: 39). Lie also notes (2013: 39) that the VESPA texts are slightly skewed towards contrastive assignments. This shows up in the concordance, where about 20 of the lines reflect a comparison (e.g. between two texts) that the student has been asked make, as illustrated by (7).

(7) The second text, *on the other hand*, is not all that engaging, just informative. (VESPA)

5.2 *When it comes to*

When it comes to is relatively widespread in VESPA, with 50 occurrences across 25 texts (23 writers), while it is absent from BAWE. This does not mean the expression is non-native; it occurs three times in CRA.[11] The syntactic function of *when*

[10] This includes the variant forms *på den andre siden* / *på den andre sida*.

[11] This corresponds to 1.8 per 100,000 words.

it comes to X is respect adjunct (as defined in Hasselgård 2010: 28, 244–248), typically specifying the circumstances under which the proposition applies. The usage of *when it comes to* in VESPA has the following characteristics:

- The bundle is typically followed by an indefinite noun phrase.
- 22 (of 50) occur in sentence-initial position (six of which are preceded by a connector); these mark the sentence topic, as in (8).
- In end position the function is typically to restrict the validity of the proposition; see (9).
- Certain sentence-final occurrences are close to postmodifying function; see (10).

(8) *When it comes to collocation and sentence structure* this student has some very strange ways of saying things ... (VESPA)

(9) But everything that is said in this short excerpt is in the present tense, which is quite normal *when it comes to this kind of literature.* (VESPA)

(10) ...the author of this text has some problems *when it comes to word formation.* (VESPA)

Seeking to explain the massive overuse of *when it comes to* in VESPA, I searched for the following near-synonyms in the corpora: *with regard(s) to, with respect to, as regards, as to, concerning, regarding* and *in terms of*. The latter turns out to be a clear favourite, outnumbering all its synonyms across the board. Figure 1 shows *when it comes to* and *in terms of* separately while all the other near-synonyms have been lumped together. The collective frequencies of such expressions do not differ significantly across the corpora; i.e. the alternative expressions make up for the overuse of *when it comes to* in VESPA.

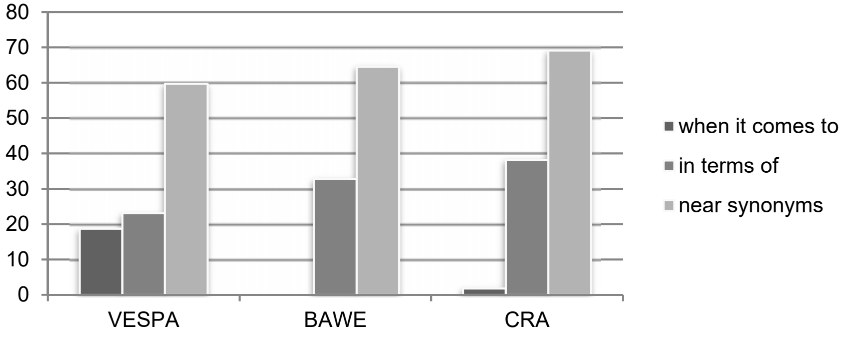


Fig. 1: *When it comes to* and its near-synonyms in three corpora.

Searches in the BNC and COCA reveal that *when it comes to* is more than twice as frequent in American as in British English (12.1 per million words in the BNC vs. 26 in COCA). It is relatively common in journalistic texts in both varieties and in speech in COCA, but infrequent in academic writing. Interestingly, the bundle has almost doubled its frequency in American English within the COCA time span (as per 2014), with 18.36 hits per million words in 1990–94 as against 35.6 in 2010–12. A possible cause of the overuse in VESPA might thus be influence from spoken American English and journalistic texts, i.e. genres that are widespread in Norwegian society through the media. However, we may note that the linguistics part of MICUSP has only one occurrence of this bundle.

Looking to the Norwegian language for a source of the phrase *when it comes to*, we find the formally similar *når det gjelder* ('when it concerns'). This expression is fairly common in all text types, including academic prose: in the KIAP corpus, the expression was found with a frequency of 20.4 per 100,000 words in linguistics articles, i.e. very close to the frequency of *when it comes to* in VESPA. See example (11).

(11) En av de sentrale forskjellene de tar opp, er nettopp forskjellen *når det gjelder* plassering av det finitte verbet i leddsetninger. (KIAP)
"One of the central differences they take up, is indeed the difference when it comes to placement of the finite verb in subordinate clauses."

The evidence presented here suggests that the overuse of *when it comes to* can be related to the Norwegian *når det gjelder*, which is functionally and formally similar. In addition, learners may find support for their use of the expression through the media. Curiously, however, in the non-fiction part of the ENPC, *når det gjelder* is more frequent in translations (from English) than in Norwegian originals. Sometimes the source of *når det gjelder* is a preposition, as shown in (12), thus perhaps suggesting that *når det gjelder* is a relatively grammaticalized expression used for relating two concepts. Norwegian learners may have transferred this to their use of *when it comes to*, as suggested by examples such as (10) above, where a more elegant wording might have been "...has some problems with word formation".

(12) There is still much to be understood about the origin of life, including the origin of the genetic code. (ENPC, CSA1)
Det er fortsatt en hel del vi ennå ikke forstår *når det gjelder* livets opprinnelse — blant annet hvordan den genetiske koden oppstod. (CSA1T)
Lit: 'There is still a whole lot we yet not understand when it concerns life's origin...'

5.3 *As we can see*

The bundle *as we can see* occurs in 27 texts in VESPA (11.3%), but not at all in BAWE or in CRA. Its use in VESPA has a clear profile:

- It is typically used in sentence-initial position (20 of the 31 instances).
- It is typically followed by a preposition (19 of the 31 instances), of which the most frequent one is *from* (15 instances).
- 11 instances are followed by either a subject NP (7) or existential *there* (4).

In sentence-initial position *as we can see* has a connective function, as illustrated by (13). It is a metadiscursive marker, typically referring to texts, tables, figures and examples discussed. The prepositional phrase following it tells readers *where* something is to be seen.

(13) *As we can see* from the above examples, *effektiv* is not used with this reference in Norwegian, and is therefore rephrased. (VESPA)

When *as we can see* is directly followed by a subject, with no specification of where to look, the location may be evident to the reader, as in (14), where the student is referring to a text s/he has been asked to analyse.

(14) There are also examples of sentence structure that clearly derives from Norwegian influence, as in sentence (13) under 'conceptual confusion'. The spelling errors may be classified as intralingual errors, and *as we can see*, there is a lot of this type of errors as well. (VESPA)

Hyland (2008) highlights, even in the title of his paper, the frequency of the metadiscourse marker *as can be seen*, i.e. the passive counterpart of *as we can see*. Table 6 shows that *as can be seen* is recurrent in both VESPA and BAWE, and interestingly it is more frequent in VESPA. However, only five VESPA writers use it, while 27 use the active phrase. It may be noted that the active phrase occurs with about equal frequencies in spoken and academic English in both the BNC and COCA, while the passive phrase is frequent only in academic prose in both varieties. The bundle favoured in VESPA is thus the more colloquial one.

Tab. 6: *As we can see* and *as can be seen* across corpora. N = raw frequencies, R = relative frequencies per 100,000 words.

	VESPA		BAWE		CRA	
	N	R	N	R	N	R
as we can see	31	11.6	0	0	0	0
as can be seen	14	5.2	5	3.0	6	3.5

As can be seen has the same usage profile in VESPA as its active counterpart: it is sentence-initial in 8 of the 14 instances; it is followed by a prepositional phrase in 11 instances and by a subject NP in the remaining three. There appears to be no difference in the discourse functions of the active and the passive; both are metadiscursive and guide the reader to tables, concordances, examples and the like, as illustrated by example (15).

(15) ...the phrase is expressing the simple future, *as can be seen* in one of the hits the PerlTCE produced. (VESPA)

The six examples of *as can be seen* in CRA are all followed by a prepositional phrase or the adverb *above*, four of the six are sentence-initial, and they signpost tables, figures and concordances. The VESPA writers have thus grasped the functions of *as can be seen* (and transferred them to *as we can see*), but use the phrase more often than native speakers.

The most closely corresponding Norwegian equivalent to *as we can see* is *som vi ser* ('as we see'). This expression is found in KIAP but is infrequent (1.1 per 100,000 words); searches for the related *som en/man ser* ('as one sees') and *som vi/en/man kan se* ('as we/one can see') add only four hits, increasing the frequency to 2.6 per 100,000. The closest counterpart in the passive voice, *som vist* ('as shown') is more frequent at 5.9 occurrences per 100,000 words, and has much the same profile as the English *as can be seen*. In any case, direct transfer from Norwegian thus cannot be the source of the overuse of *as we can see* observed in VESPA.[12]

A possible explanation might instead be found in the general tendency of learners towards overuse of metadiscourse (see Ädel 2006, Hasselgård 2016) and

[12] Lee and Chen (2009: 289) identify *we can see* as one of the collocations favoured by Chinese learners of English. In Lee and Chen's material too, the bundle is typically used "to refer to or explain tables or figures, and to organize the discussion" (2009: 289).

writer/reader visibility features (see Paquot, Hasselgård & Ebeling 2013). *As we can see* gives the writer the opportunity to be visible, involve the reader and organize the text at the same time, thus serving three of the functions often attributed to learner discourse.

5.4 *As a result of*

The last bundle to be discussed here is one that is underused in VESPA compared to BAWE, with only eight hits (3.4 per 100,000) occurring in eight texts (3.4%) by eight different writers. The bundle is frequent in BAWE with 26 hits (13.7 per 100,000) distributed over 13 texts (17.1%) and eight writers. CRA contains eight instances of the bundle (4.7 per 100,000), which actually does not constitute a significant difference from VESPA (LL = 0.81). The use of the bundle in VESPA has the following characteristics:

- In three out of the eight instances, *as a result of* is followed by the pronoun/determiner *this* (example 16), thus marking cohesion with the preceding context.
- Three of the other instances echo a wording in the students' textbook, see (17).
- To a greater extent than the bundles discussed above, *as a result of* occurs in sentences that contain errors, see (17) and (18).

(16) The sentences in text 2 are *as a result of* this shorter,... (VESPA)
(17) Johansson further explain them *as a result of* overgeneralisation from what the learner of language already has learnt... (VESPA)
(18) *As a result of* this the highly differ in style and form. (VESPA)

In BAWE *as a result of* typically precedes a complex noun phrase, as in (19), sometimes involving a nominalized process, as in (20).

(19) Once again, *as a result of* the more informal nature of the group interaction, ellipsis was commonplace. (BAWE)
(20) *As a result of* this examination of three types of instruction, each based on a different theory of language learning, I can now view my Persian learning in a more informed light. (BAWE)

The usage found in CRA closely resembles that of BAWE: the complement of *of* is most typically a complex noun phrase, as illustrated by (21).

(21) ...and decisions are made *as a result of* multiple identifications with value premises at an individual level, ... (CRA)

As a result of has a literal counterpart in Norwegian: *som et resultat av* (with the variant *som resultat av*). However, translations in ENPC non-fiction show that the two expressions do not often correspond: *som (et) resultat av* occurs only once as a translation of *as a result of* and twice as its source. The related *som følge av* (‘as consequence of’) is a recurrent source (5 out of 21 instances), and *på grunn av* (lit. ‘on reason of’ = ‘because of’) is a recurrent translation (3 out of 8). Both *som (et) resultat av* and *som følge av* were found in KIAP. Interestingly, *som et resultat av* is also used in a more literal sense, as illustrated by (22).

(22) ...at språkendringene blir forklart *som et resultat av* at en folkegruppe er blitt rammet av pest, krig eller andre sosiale tragedier... (KIAP)
‘that the language changes are explained as a result of [the fact] that a population has been hit by plague, war or other social tragedies...’ [My translation]

The contrastive observations may suggest that Norwegian learners of English fail to use *as a result of* idiomatically partly because of the low degree of correspondence between this bundle and its most similar Norwegian counterpart *som (et) resultat av*, and partly because the Norwegian expression seems to be less grammaticalized than the English one. But since *som resultat/følge av* is not infrequent in Norwegian, this cannot be the whole story. The VESPA contributors do write about causal relations: there is a much more frequent use of the conjunction *because* than in BAWE with 179.2 vs. 83 occurrences per 100,000 words, though it occurs in similar proportions of the texts (65% vs 68%). A more detailed investigation of causal expressions in learner language is needed to map the range of lexicogrammatical resources employed by the advanced learners. At this stage we may hypothesize that the underuse of *as a result of* has intralingual rather than interlingual causes: previous studies of Norwegian-based learner English have indicated that learners struggle with (or avoid) complex noun phrases (e.g. Hasselgård 2012, who examined the nouns *fact, issue, question* and *problem* in learner English); thus it is possible that *as a result of* represents a level of complexity that the learners are not prepared to handle.

5.5 Summary of case studies

The case studies presented in this section have illustrated some differences between learner and native speaker phraseology. The usage profiles worked out on the basis of concordances show that the learners do not always use bundles in the same contexts and with the same discourse functions as native speakers. *On the other hand, when it comes to* and *as we can see* are overused by the learners in terms of frequencies across the corpora as well as text dispersion. *As a result of* is underused. Explanations for both overuse and underuse can be sought in comparisons with the learners' L1 (in this case Norwegian) and in general reference corpora of spoken and written English. It has been found here that the three overused items have more similar corresponding expressions in Norwegian than the underused one. Comparison with expert writing in the same discipline is important. For example, in selecting an underused bundle for further examination I realized that most of the bundles that are underused compared to BAWE (Tables 3 and 5) are not underused if compared to CRA instead. The overuse, however, is at least as strong.[13]

The bundles *on the other hand* and *when it comes to* display two important characteristics of lexical/phraseological teddy bears: they are much more frequent in English L2 than in English L1, and they seem to have generalized their meanings and discourse functions by being used in contexts where native speakers prefer other expressions. The third overused expression, *as we can see*, does not show the same pattern of generalization, and can probably be ascribed to the learners' general leaning towards a colloquial style. The preference for a colloquial style may partly explain the underuse of *as a result of*, although limited proficiency may also cause learners to avoid this complex construction.

6 Conclusion

The present study has looked into four-word bundles only. Given the relatively short lists of bundles frequent enough to reveal patterns of use, it is unlikely that a study of longer bundles will be very fruitful. However, the inclusion of three-

[13] Unfortunately the CRA interface does not show text dispersion, so overuse and underuse can only be calculated from frequency of occurrence. Nor does the interface allow the bottom-up extraction of bundles (or downloading of raw texts). It is, however, highly likely that other underused bundles would have been identified in comparisons between VESPA and CRA.

word bundles may be able to complete the picture (cf. Lie 2013; Ebeling & Hasselgård 2015a). Furthermore, as many of the studies referred to in Section 2 point out, lexical bundles differ across disciplines. A natural next step would thus be to make a similar investigation of the other disciplines available, as VESPA now also contains literature and business texts.

Despite its limitations, this study has shown that frequency patterns of recurrent lexical bundles differ between learners and native speakers: the most frequent bundles are *more* frequent in learner English, but frequencies drop less sharply in L1 English. This indicates that the learners have clear favourites that they “clutch for” and “feel safe with” (Hasselgren 1994: 237), so we may justifiably speak of ‘phraseological teddy bears’.

The study of text dispersion gave a different rank frequency of the bundles than the one based on frequencies of occurrence. The most common bundles turned out to occur in a greater proportion of the texts in L1 English; learners are thus less uniform in their use of most of the frequent bundles. Some learners appear to be more fond of their phraseological teddy bears than others. This suggests that text dispersion may be a better indicator than frequencies per 100,000 words of over- and underuse of lexical bundles.

The qualitative studies of four selected bundles corroborate Ädel and Erman’s (2012) prediction that learners and native speakers may not be using bundles in the same way. The case studies were performed only on bundles with significantly different frequencies between VESPA and BAWE, but should in principle also be carried out for other bundles, as similar frequencies do not automatically mean similar usage. By the same token, differences in frequency need not imply that a bundle is used incorrectly by the learners. A methodological feature of this study consists in cross-checks with parallel corpora and corpora of expert academic writing. This is believed to be indispensable in addressing potential sources of transfer as well as the issue of discrepancies between novice and expert writing. Furthermore, the qualitative study of individual bundles in terms of profiles of usage, as carried out in Section 5, seems to be a fruitful way of exploring divergences between native and non-native style.

This type of qualitative analysis is certainly required before any pedagogical recommendations can be made concerning the use of lexical bundles and phraseological teddy bears. We need to know “what they are, how and why we use them, how they affect our discourse and, hopefully, how we might be persuaded to part with them” (Hasselgren 1994: 237). Rather than just being told to use *when it comes to* less and *as a result of* more, learners should be made aware of the appropriate contexts and functions of the bundles. It is also important to compare L1 novice writing with expert writing before trying to change L2 behaviour since the professional writers are more likely to represent a learning target. Both quantitative and qualitative analyses are needed to tease out differences between native and non-native phraseology. It seems, however, that pedagogical applications should be derived from patterns of underuse: rather than depriving the learners of their phraseological teddy bears we should give them some new toys.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam, Netherlands: John Benjamins.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native speakers and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 101–122). Oxford, UK: Oxford University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). 'If you look at…': Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London, UK: Longman.

Byrd, P., & Coxhead, A. (2010). *On the other hand*: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*, 31–64.

Chen, Y-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 30–49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–323.

Ebeling, S.O., & Hasselgård, H. (2015a). Learners' and native speakers' use of recurrent word-combinations across disciplines. *Bergen Language and Linguistics Studies, 6*(1), 87–106.

Ebeling, S.O., & Hasselgård, H. (2015b). Phraseology in learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 207–230). Cambridge, UK: Cambridge University Press.

Ebeling, S.O., & Heuboeck, A. (2007). Encoding document information in a corpus of student writing: The British Academic Written English corpus. *Corpora, 2*(2), 241–256.

Ellis, N.C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44.

Fløttum, K., Dahl, T., & Kinn, T. (2006). *Academic Voices*. Amsterdam, Netherlands: John Benjamins.

Fløttum, K., Dahl, T., Didriksen, A. A., & Gjesdal, A. M. (2013). KIAP–reflections on a complex corpus. *Bergen Language and Linguistics Studies*, *3*(1), 137–150.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing. In Cowie, A.P. (Ed.), *Phraseology: Theory, Analysis and Applications* (pp. 145–160). Oxford, UK: Oxford University Press.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *52*(3), 229–252.

Hasselgård, H. (2010). *Adjunct Adverbials in English*. Cambridge, UK: Cambridge University Press.

Hasselgård, H. (2012). *Facts, ideas, questions, problems*, and *issues* in advanced learners' English. *Nordic Journal of English Studies*, *11*(1), 22–54.

Hasselgård, H. (2016). Discourse-organizing metadiscourse in novice academic English. In M.J. López-Couso, B. Méndez-Naya, P. Núñez-Pertejo, & I.M. Palacios-Martínez (Eds.), *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora* (pp. 106–131). Leiden, Netherlands: Brill | Rodopi.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237–259.

Heuboeck, A., Holmes, J., & Nesi, H. (2008). The BAWE Corpus Manual. Available at http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf (accessed May 2016, last accessed October 2018).

Hunston, S. (2008). Starting with the small words. Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics*, *13*(3), 271–295.

Hyland, K. (2008). As can be seen. Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21.

Lee, D. Y. W., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, *18*(4), 281–296.

Levenston, E.A., & Blum, S. (1977). Aspects of lexical simplification in the speech and writing of advanced adult learners. In S.P. Corder & E. Roulet (Eds.), *The Notion of Simplification, Interlanguage and Pidgins and their Relations to SL Pedagogy* (pp. 51–71). Neuchâtel, Switzerland: Actes du 5ème colloque de Linguistique Appliquée de Neuchâtel.

Lie, J. (2013). "The fact that the majority seems to be…" A corpus-based investigation of non-native academic English. Master's dissertation, University of Oslo, Norway. Retrieved from http://urn.nb.no/URN:NBN:no-46382 (last accessed October 2018).

Moon, R. (1998). *Fixed Expressions and Idioms in English. A Corpus-based Approach*. Oxford, UK: Clarendon.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam, Netherlands: John Benjamins.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics*, *18*(3), 391–417.

Paquot, M., Ebeling, S.O., Heuboeck, A., & Valentin, L. (2010). The VESPA tagging manual. Université catholique de Louvain and University of Oslo. Unpublished manuscript.

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Paquot, M., Hasselgård, H., & Ebeling, S.O. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead* (pp. 377–287). Louvain, Belgium: Presses Universitaires de Louvain.

Pawley, A. & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In S.C. Richards & R.W. Schmidt (Eds.), *Language and Communication* (pp. 191–226). London, UK: Longman.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, *14*, 84–94.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer* (pp. 41–52). London, UK: Longman.

Salazar, D. (2014). *Lexical Bundles in Native and Non-native Scientific Writing*. Amsterdam, Netherlands: John Benjamins.

Scott, M. (2012). WordSmith Tools version 6. Stroud, UK: Lexical Analysis Software.

Simpson-Vlach, R., & Ellis, N.C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language*, *10*(1), 61–104.

Corpora

BAWE (British Academic Written English Corpus): http://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/ (accessed June 2014, last accessed October 2018).

VESPA (Varieties of English for Specific Purposes dAtabase): https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html, http://www.hf.uio.no/ilos/english/services/vespa/ (accessed June 2014, last accessed October 2018).

CRA (Corpus of Research Articles): http://rcpce.engl.polyu.edu.hk/RACorpus/ (accessed June 2014, last accessed October 2018).

BNC (British National Corpus): http://www.natcorp.ox.ac.uk/ (accessed June 2014, last accessed October 2018).

COCA (Corpus of Contemporary American English): http://corpus.byu.edu/coca/ (accessed June 2014, last accessed October 2018).

ENPC (English–Norwegian Parallel Corpus): http://www.hf.uio.no/ilos/english/services/omc/enpc/ (accessed June 2014, last accessed October 2018).

KIAP (Cultural Identity in Academic Prose): https://www.uib.no/fremmedsprak/23107/kiap-korpuset (accessed June 2014, last accessed October 2018).

MICUSP (Michigan Corpus of Upper-Level Student Papers) http://micusp.elicorpora.info/ (accessed June 2014, last accessed October 2018).

Rolf Kreyer

“Dear ~~Man~~ ~~men~~ ~~and women~~ madam, dear ~~xxx~~ sir”

What we can learn from revisions in authentic learner texts

Abstract: The present chapter explores to what extent methods of writing process analysis can be fruitfully applied to learner corpus data. An analysis of 598 instances of revision in learner data shows that the majority is of a conceptual nature, i.e. involves changes of content. Formal revisions, i.e. revisions demanded by the target language system and thus particularly indicative of the state of the interlanguage system, come second. Within this group, grammatical, lexical and orthographic/typo revisions are far more frequent than revisions that concern questions of idiomaticity or textual cohesion and coherence. This is interpreted as a lack of awareness of the latter aspects on the part of the learner possibly due to a lesser prominence in the EFL curriculum. In addition, it is found that in more than three quarters of all cases of formal revisions, the text is improved. It follows that the quality of the final product may mask problem areas of the learner that only become apparent if we take the writing process into consideration.

1 Introduction

Writing process analysis is a research paradigm that was started off by the influential work of Emig (1971) and initiated a change of perspective on text: text was no longer seen as a mere product but as a process, a shift of consciousness that was described as the “most enduring of her [Emig’s] contributions” by Voss (1983: 278; see Raimes 1991 for a discussion of different approaches to writing from 1966–1991), and still prevails today. The aim of this approach is to tap into the cognitive processes that are going on in the writer’s mind during writing. The approach has been fruitfully applied to areas as diverse as creative writing or other kinds of professional writing, translation or subtitling, studies in the development of writing in L1 as well as research in the development of a foreign language (cf. Van Waes et al. 2012: 507). The present chapter falls into the latter kind of

Rolf Kreyer, University of Marburg, kreyer@uni-marburg.de

https://doi.org/10.1515/9783110489071-014

research, trying to combine writing process analysis, more specifically revision analysis, with learner corpus data.

Writing process research, as Park and Kinginger (2010: 31) state, so far, "has relied upon two major data sources: (a) retrospective accounts from participants [...] and (b) audio/video data collected in real time [...]". While both kinds of approaches have their advantages (and their disadvantages), a major drawback is the fact that, because of the complexity of the procedures, the number of informants is usually rather low (Emig 1971, for instance, analysed the writing processes of eight senior students). The present chapter explores to what extent writing process analysis might benefit from the significantly larger amounts of data accessible through learner corpora and how the consciousness that characterises writing process analysis could give rise to new questions, new approaches and new answers in learner corpus linguistics.

From a learner corpus perspective "[t]here is no doubt that the efficiency of EFL tools could be improved if materials designers had access [...] to authentic learner data [...] highlighting what is difficult for learners [...]" (Granger 1998a: 7). However, one drawback of most (learner) corpora is that they provide the researcher with the finished product, thereby ignoring the process that led to this product. This becomes particularly relevant if we conceive of the learner texts as a product of interlanguage in the sense of Selinker (1972), i.e. as a language system in its own right which, however, is highly unstable as a learner usually is on his/her way to the target language system (see also Corder's 1981: 18 term 'transitional dialect'). Given this instability of the system, a focus on the final product can lead to a distorted view of what the learner can or cannot do. In example (1), for instance, the finished product masks the fact that the learner still has problems with the verb *to accept*, most probably due to interference from the German language, *akzeptieren*, and false analogy based on pairs like *nominieren – nominate* or *regulieren – regulate*.

(1) Mr. Brock
doesn't **~~acceptate~~** accept all technology
ande murder them.

Conversely, the finished product may be an erroneously revised version of a correct sentence, as in (2).

(2) [...] but that there **~~are~~** parents are who don't want it, is such a shame.

The final version of the sentence leads us to assume that the student is not aware of word order in the *that*-clause; an assumption that is proven wrong if we have access to data that gives us insight into the writing process: the student first uses the correct word order, and falls back on the German pattern in the revision. The **M**arburg corpus of **I**ntermediate **L**earner **E**nglish (MILE) provides us with such data.

The present chapter, then, aims to demonstrate that the MILE can be used to provide a quantitative basis for writing process analysis: by analysing the revisions in the corpus, in particular those of a formal-conventional type (see Section 5), we can get a clearer idea of interlanguage competence.

2 Learner Corpora and the MILE

For the last two or three decades, learner corpora have become increasingly important for the study of the interlanguage of foreign language learners, particularly learners of English. However, despite a large and quickly increasing number of learner corpora we are still confronted with a paucity of data from learners of English on a beginning to intermediate level. Barlow's observation from 2005 is still valid after more than 10 years: "[m]ost of the existing learner corpora are based on the writing of fairly advanced language learners" (2005: 357). Corpora that do represent beginner to intermediate learners are either very difficult to access (e.g. the large commercial corpora compiled by publishers like Cambridge and Longman) or are rather small and/or not available to electronic analysis.[1] On the whole, it seems fair to conclude that the representation of intermediate learners of English has been neglected so far. This is most probably due to a number of serious bureaucratic and data-protection problems that come along with the compilation of data from adolescents. In addition, hand-written exams are far less easily digitised than university prose, which is usually available in digital format. Although understandable, the lack of corpora representing younger learners is deplorable (see Kreyer 2015).

Similarly, second language acquisition research in general and corpus-based research into SLA in particular suffer from the scarcity of longitudinal data: "a considerable percentage of SLA research, if carefully examined in terms of its focus and data, does not actually address the process of acquisition per se, as it relies on and investigates cross-sections of L2 language use" (Hasko 2013: 2; also

[1] Examples include the Interactive Spoken Language Education (ISLE; Atwell, Howarth, & Souter 2003), the Flensburg English Classroom Corpus (FLECC; Jäkel 2010) and the International Corpus of Crosslinguistic Interlanguage (ICCI; Tono 2012).

see Ortega & Iberri-Shea 2005 for a similar view). Although corpus linguists have repeatedly called for longitudinal corpora (Barlow 2005; Granger 2002, 2008 and 2009; Leech 1998; Ortega & Byrnes 2008), these calls seem to have gone unheard so far. A recent survey observes a "paucity of longitudinal corpora" (Hasko 2013: 3) and states that "much work remains to be done" (Hasko 2013: 2; see Kreyer 2015).

To sum up so far, there are currently two considerable gaps in the learner corpus landscape, namely (1) the scarcity of material representing beginning or intermediate (German) learners of English, and (2) the even more drastic lack of real longitudinal learner data. The MILE is an attempt to fill these gaps.

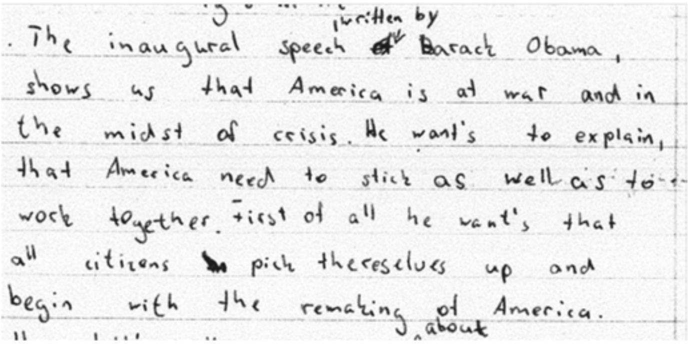
The MILE is currently being compiled at the University of Marburg. It aims at creating a database of written learner English from grades 9 to 12 of a German secondary school. The compilation started in 2011 with 149 pupils from grade 9 in the school year 2011/12 in one German school. Pupils in grade 9 usually are 14 to 15 years of age and have already had at least six years of English in school. At the time of writing (November 2014), we are collecting data from 89 of these 149 pupils, since 60 pupils have left the original cohort. Table 1 provides an overview of the estimated number of words that the finished MILE will contain (pending approval of some pupils and parents to use their data).

Tab. 1: Estimated number of words in the MILE.

grade	# of pupils	# of words
9	149	108,000
10	135	217,000
11	93	260,000
12	89	182,000
		~767,000

As can be seen, the corpus is designed as a longitudinal database that documents the progress of learners in their final years of secondary education. The data mostly consists of written material from text production tasks in official exams. In some cases, these timed exams have been replaced (by the examiner) by untimed forms of assessment, such as book reports (see Kreyer 2015 for details).

Originally, it was intended to provide PDFs of the hand-written exams together with the digitised data, because the writing process itself, as documented in revisions in the text, can reveal a lot about the interlanguage of the writer. However, since this idea met with strong opposition from data protection officers, ‘quasi-facsimiles’ of the original hand-written documents had to be created. Figure 1 below is an example of a hand-written text, Figure 2 shows the corresponding quasi-facsimile.



The inaugural speech ~~of~~ ,written by Barack Obama, shows us that America is at war and in the midst of crisis. He want's to explain, that America need to stick as well as to work together. first of all he want's that all citizens ~~xx~~ pick thereselves up and begin with the remaking of America.

about

Fig. 1: An example of a hand-written text collected in the MILE project.

The inaugural speech ~~of~~ , written by Barack Obama,
shows us that America is at war and in
the midst of crisis. He want’s to explain,
that America need to stick as well as to
work together. First of all he want’s that
all citizens ~~xx~~ pick thereselves up and
begin with the remaking of America.

Fig. 2: A quasi-facsimile of the example shown in Figure 1.

Figures 1 and 2 illustrate how valuable the editing process is with regard to studying interlanguage: in the above case, for instance, the student seems to have a problem with the use of the correct preposition, which s/he solves by using *written by*. In the final, ‘authorised’ version of the text, nothing hints at that problem (see above).

The information captured in the quasi-facsimile is also stored in the text-internal markup shown under (3).

(3) The inaugural speech <d>of</d> <ail>, written by</ail> Barack Obama, <lb></lb> shows us that America is at war and in <lb></lb> the midst of crisis. He want’s to explain, <lb></lb> that America need to stick as well as to <lb></lb> work together. First of all he want’s that <lb></lb> all citizens

> <d><ur>2 letters</ur></d> pick thereselves up and <lb></lb>begin with the remaking of America.
> Key: d – deleted; ail – added in line; lb – line break; ur – unreadable.

Fortunately, what started as a workaround has given us access to a whole new set of data, since the text-internal markup shown in (3) can be identified by the computer and become an object of linguistic enquiry and analysis. That is, all kinds of revision made in the original documents can now be analysed with corpus linguistic methods. This provides us with an additional window onto interlanguage. While error analysis, with its focus on deviance in the final product, tells us something about problems that students are not yet aware of (or simply have no idea at all of how to correct), revision analysis, with its focus on the process, can provide us with evidence of difficulties which students are aware of. Error analysis and revision analysis, thus, complement each other nicely and together provide us with a more precise view on (emerging) learner language competence.

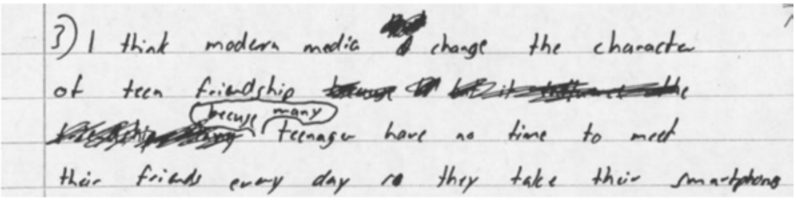
3 The data

The data at hand contains digitised materials from grade 9 to grade 11 as can be seen in Table 2.

Tab. 2: The number of words per grade in the present data set (numbers rounded). The number for grade 11 is different from that shown in Table 1 since not all of the material from grade 11 has been available for analysis.

grade	# of words
9	108,000
10	217,000
11	152,000
total	477,000

To identify revisions, all instances of deletions were searched for in the corpus by looking for the tag '<d>', i.e. for all revisions that involve deletions of text material. Figure 3 is an example of a heavily revised text, the corresponding marked-up version is provided in (4).



3) I think modern media [illegible] change the characte
of teen friendship ~~becaus~~ ~~b~~ [illegible] ~~the~~
[illegible] becuse many teenage have no time to meet
their friends every day so they take their smartphone

Fig. 3: An example of a heavily revised text.

(4) I think modern media <d><ur>1 letter</ur></d> <ail><d><ur>1 word</ur></d></ail> change the character <lb></lb> of teen friendship <d>because</d> <d>b</d> <d>but it <ur>1 word</ur> the</d> <lb></lb> <d>friendship of any</d> <ail>becuse many</ail> teenager have no time to meet <lb></lb> their friends every day so they take their smartphones

As can be seen from the above, much of the deleted material is unreadable in Figure 3. Fortunately, this is not representative of the material in general: of the 15,302 deletions found in the material, roughly 20% is not readable, i.e. we are left with 12,087 instances of readable deletions. From these, a random sample of 598 were taken for the present analysis.

4 Classifying Revisions

Online revisions in writing can, of course, be described in a variety of ways. Abdel Latif (2008: 37) describes in an exemplary fashion three taxonomies of revisions (New 1999; Stevenson, Schoonen, & de Glopper 2006; Van Waes & Schellens 2003). There is no need to go into detail here, but it is not surprising that a vast array of (overlapping) aspects can be taken into consideration, such as the following:

- the linguistic level on which the revision takes place, ranging from letter to paragraph and layout
- the motivation for the revision, e.g. correcting errors or changing content
- the type of error that has been corrected, e.g. punctuation, spelling, or grammar errors
- the type of revision, e.g. adding, deleting or substituting material
- the type of formal change, e.g. abbreviation, contraction, punctuation etc.

- the location of the revision, e.g. in-line, on top of the original text, on the margin of the page, etc.
- ...

The taxonomy applied in the present study is a modified form of Lindgren (2005). Lindgren works with keystroke logging, a technology that allows to specify when exactly in the writing process the revision occurred. Thus, a ‘contextual revision’, i.e. a revision that is made to an already existing text, can be distinguished from a ‘pre-contextual revision’ that occurs at the position in the text that is currently being written.

The present data, for obvious reasons, does not provide that kind of information, which is why the classification of the data at hand begins with the second level of Lindgren’s taxonomy, namely the distinction between formal and conceptual revision. The former is defined in negation of the latter: “[...] Form, include [sic!] revisions that do not affect the content of the text [... whereas] [c]onceptual revisions affect the content of the text” (Lindgren 2005: 20). With regard to form a further distinction is made between conventional and optional revisions i.e. whether a revision is required by the language system or not. In addition, the taxonomy captures whether a formal revision “corrects or creates a mistake or does neither” (Lindgren 2005: 20), i.e. formal revisions are categorised as ‘correct’, ‘erroneous’ and ‘neutral’, respectively. I will only apply this distinction to conventional but not to optional revisions. The former are the most interesting for the present study as they show the learners’ awareness of incorrect language use and, therefore, provide a particularly informative view on the developing interlanguage. Lindgren (2005) further distinguishes between conceptual revisions for stylistic or audience-design reasons – a distinction that will be ignored in the present chapter. The basic taxonomy is shown in Figure 4.

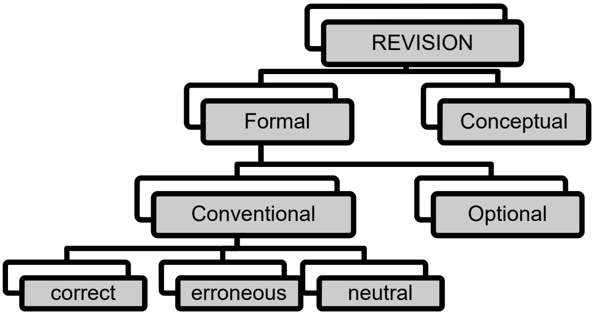


Fig. 4: The basic taxonomy of revisions used in the present study.

Note that the distinction between correct and erroneous revisions is applied locally. In example (5) below, the revision from *did* to *dit* is incorrect. Whether or not the final version *they **make** some social experience* is correct or not is irrelevant.

(5) [...] And after
they **~~did~~** **~~dit~~** make some social experience and have
understood that they can do something
meaningful in their life [...]

An exception to this are cases where a change at one point in the text leads to a number of concomitant changes, illustrated in example (6).

(6) When **~~you~~** **~~she~~** **~~are~~** **~~is~~** **you are** arguing with **~~your~~** **~~her~~** **your**
parents **~~you~~** **~~she~~** **you** **~~should~~** **have to** calm down to find
good arguments and logic arguments.

It makes sense to assume a sequence of changes which does not coincide with the sequence of the words: all the words in superscript appear to be the result of one large revision which results from an uncertainty as to whether *you* or *she* should be the subject of the sentence (in the end, the student chooses *you*). Consequently, the changes in the person of the verb are not interpreted as local grammar-related changes but as global content-related changes since they result from a content-related change in the beginning of (6). Since the focus of the present study is on the analysis of interlanguage, particular emphasis is placed on conventional revisions: what can these revisions tell us about the L2 competence of German learners of English?

5 Revisions – a window onto L2 competence of German learners of English

As described above, the data at hand comprises 598 tokens of revision. A first breakdown of these into five major categories is given in Table 3.

Tab. 3: A breakdown of revisions into major categories.

kind of revision	frequency	
conceptual	273	45.7%
formal_conventional	218	36.5%
formal_optional	11	1.8%
no change	53	8.9%
unclear	43	7.2%
total	598	100%

Conceptual revisions (accounting for 46% of the data), as described above, are concerned with the content and style rather than with the linguistic form. Revisions of that kind may concern semantic specifications (7), lexical variation (8) and (9), stylistic appropriateness (10) or the rewriting of longer passages (11).

(7) The film about his dead is ~~26~~ circa 26 second and was published
in 1975.

(8) Sowell probably thjinks that most people only see the advantages
of technology and the criticers are in the minority. Mr. Brock is in
the same position so there's another common ground with the
short story.
~~Another~~ Moreover Sowell says that technology has advantages
although it can have many bad consequences. Mr. Brock is at the
same opinion.

(9) [...] Because you are not restricted by something
which ~~rest~~ limits you in other countrys.

(10) Todd is a very shy ~~guy~~ boy, but with Mr. Keatings
he find [...]

(11) But the most important point is that ~~youth~~ teenager doesn’t interests in the
politic [...]

(12) I had make an "Schüleraustausch" to England ~~before~~ for one year
– So when I can English very good, I have good chance of a job
there and I see my boy-friend all day.

Some of the examples above make clear that at times the analysis involves a certain degree of interpretation. (9) is analysed as a conceptual change instigated by the need for lexical variation because it makes sense to assume that the writer originally wanted to use *restricts* but then decides to use *limits* because s/he has already used *restricted* in the same sentence. In addition, for an estimated 10 %

of conceptual revisions, it was not easy to decide whether we are dealing with a conceptual revision or a formal one, as in example (11). This revision might be due to problems the pupil has with the use of *youth*, e.g. whether to use a definite article or not. Similarly, example (12) might indicate that the writer has problems with particular conjunctions. However, it is also possible that both revisions are driven by reasons relating to content or style. Since the focus of this chapter is on interlanguage and, hence, on formal revisions, I decided to take a rather conservative approach in the analysis of my data: if the deleted material was not incorrect and the revision could be interpreted as conceptual (as in (11) and (12)), it was categorised as such to ensure (as much as possible) that the items that end up in the formal revision 'bin' really tell us something about language competence and not about content-related changes.

The second largest category, with a total of 229 (218+11; 38.3%) tokens, is that of formal revisions. Among these, optional revisions, i.e. those that make unnecessary changes to an area of grammar or lexis, are fairly rare (11 tokens; 1.8%). Two examples are given below. In (13) the writer chooses to insert a subordinating conjunction, in (14) s/he decides on the relative pronoun *who* instead of *which*.

(13) They thought ~~**God**~~ xx that God told them to go to [...].
(14) the norther states, ~~**which**~~ who were
against the slavery [...]

Conventional revisions, i.e. those that are demanded by the English language system, in contrast, are far more frequent, namely 218 (36.5%). These kinds of revisions can be correct, erroneous or neutral, as shown in the examples below.

(15) [...] and if Mr Keating ~~**woudl**~~ wouldn't be, the Neil wouldn't
commit suicide.
(16) [...] that Mr Neck will be fired and that we ~~**can always**~~ can say
always our opinion.
(17) And this humour ~~g~~ gets ~~**thou**~~ throg the hold chapter sometimes
more sarcastic [...].

A category that has not been discussed above is the category 'No change'. This category encompasses those tokens that are not actual revisions since the writer deletes a portion of the text only to replace it by the same string. The strings are either morphemes, whole words or substrings of words which are not a linguistic unit, as shown in examples (18) to (20), respectively. Note that, even though such 'revisions' do not lead to any changes, they still may hint at areas that are problematic for the writer.

(18) [...] because in some schools were the meals very **~~un~~un**delicious.
(19) [...] it seems like no one **~~wants~~ wants** any more immigrants - [...].
(20) [...] Eschenbach is about the ~~xxx~~ ~~xxx~~ ~~statement~~ **~~opin~~ opin**ion that
young people learn in their youth [...].

Finally, there was need for a class of unclear cases (43 tokens, 7.2%), which even by inclusion of the larger context, could not be clearly assigned to one of the above categories. In (21) for instance, it is not clear, whether the writer became aware after writing *it* that the sentence already has a subject (which then would be a conventional change), or whether s/he wanted to change the subject from *He* to *it* (a conceptual change), or whether s/he originally wanted to use *iterate* as a verb but then changed this to *says*. In (22) we do not know whether the revision is a case of avoidance because the writer is not quite sure about the use of the indefinite article (*a* as opposed to *an*), which would then be an instance of conventional revisions or whether s/he changed her mind about what s/he wanted to say, a conceptual revision.

(21) He **~~it~~** says that without immigrants, he would not
stands here.
(22) Although he is **~~an tired~~** sad, because his
family hasn’t got much money.

In some cases, the deleted material is only one or two letters, which usually makes it impossible to decide on the status of the revision, as in (23).

(23) Immigrants are have the right for education in this country **~~t~~** who are live
~~there~~ ther.

Sometimes, several interpretations are possible, but since there is no conclusive evidence for one option no decision is made, as in (24). The deleted letter ‘h’ in the last-but-one line might be interpreted as an orthographic problem concerning the spelling of the relative pronoun *who* or a conceptual revision as in *When a student works for companies he* .... This and similar cases were categorised as ‘unclear’.

(24) There are also students who ~~he make~~ $^{\text{reduce}}$ their
own carbon footprint with the gap year.
But this is not all, a "gapper" can ~~not~~
even reduce carbon footprints of other
humans.

When student works for companies ~~**h**~~ who
tr~~ies~~ y s to avoid CO2 emissions [...]

5.1 A closer look at interlanguage – zooming in on formal-conventional changes

One of the aims of the present chapter is to explore to what extent the analysis of revisions can tell us something about the emerging interlanguage system. Therefore, it makes sense to look at those instances of change that are demanded by the system of the L2, i.e. formal changes that are of a conventional kind. A closer look at the 218 tokens reveals that formal-conventional revisions concern many different areas of the language system. All in all, seven categories can be distinguished. The first major category is that of ‘grammar’, such as revisions with regard to tense, word order or the pronoun system.

(25) they ~~**was**~~ were
not allowed to go to ~~hos~~ the hospitals for white
Americans.
(26) But what ~~**they**~~ would ~~**would**~~ they do?
(27) I hope you will
never going to be oneof this high sorciety cooks ~~**who**~~ which only
~~whan~~ wanted to be famous.

The category ‘lexis’ captures those changes that concern choice of words without ‘major’ changes in meaning (these would be categorised as ‘conceptual’; see above).

(28) ~~Chi~~ Christmas is the ~~**Party**~~ festival of love.
(29) Of course you could also drive less car and
~~**byci**~~ cycle instead or catching a train

Revisions of lexis are distinguished from revisions regarding idiomaticity. These mostly concern unusual collocations and larger instances of prefabricated language.

(30) And the other messages
like ~~xxx~~ knowing ~~**a strange country**~~ other cultures are the
things, ~~that~~ which make the gap year so great

(31) **~~From the experience I know adult~~**
The adults think that teenagers in the age of 15 or 16 don't have
a own political opinion

In example (30), for instance, the writer uses the collocation *a strange country* instead of *a foreign country*. Both *strange* and *foreign* would be possible translations for the German word *fremd* as in *ein fremdes Land*. However, the collocation in this example is not idiomatic. Example (31) is more complex but can also be understood on the background of the German language. Most probably, the writer started off with a word-by-word translation of German *Aus der Erfahrung weiß ich, dass Erwachsene* ..., which led to an unidiomatic wording in the English text, where the idiomatic counterpart would be something like "From my own experience I know that adults ...".

The category 'text' contains tokens that relate to questions of reference and cohesion. For example, the (in)correctness of the revision of the article in ***~~the~~*** *a* *message* in (32) can only be established if we look at the larger context. Since the previous sentence mentions a text message, the revision is erroneous.

(32) The text passage from Nick Hornby's
"Slam" written 2008 was about a boy named Sam,
who gets a text message ~~of~~ from his ex-
girlfriend Alicia (on his birthday) witch
makes him think that she is pregnant.
As Sam gets **~~the~~** a message of from his exgirlfriend Alicia ~~he is kind~~
he is kind of frightened.

The categories 'orthography' and 'typo' are similar in that both relate to the right spelling of lexical items. Accordingly, they are not easily distinguished. Typical cases of orthographic revisions are those where we can see that the writer tries out different versions. In (33), the writer is not quite sure about the 'e' in *zone* and tries out an alternative version *zon*, which s/he then discards for the correct writing *zone*. A similar example is (34), which seems to indicate problems with the orthography of *TV program* and *TV series*. By settling for *TV show* s/he avoids the problem. (35) shows that the writer has problems with the correct spelling of *because*. In this case, we might even argue that his/her insecurity leads the writer to rearrange the whole sentence. (35) would thus be another example of avoidance.

(33) He gives teachers the power to make every school to a junk-food-free
~~zone~~. **~~zon~~ zone**

(34) [...] Two and a half Men thats a great TV ~~**prog ser**~~ show. ☺

(35) First Ban Ki-moon, UN secretary general,
has the opinion, that the developed countries
have to bear the most responsibility for
global warming. ~~, **because beceu because**~~
~~the factories of these~~
~~He thinks that the factories in these countries~~
He thinks the reason for the climate change
are the factories, which emit ...

As said above, orthographic revisions are not always easy to distinguish from typo revisions. One criterion that was employed is whether the deleted part is a full word or not; in the second case the token was usually classified as a typo revision. The distinction is illustrated in (36): the first revision would be classified as ‘orthographic’ since the deleted material is a whole word, and it makes sense to assume a problem relating to orthography because nouns are capitalised in German. With the second revision, in contrast, it makes sense to assume that the writer ‘forgot’ to include the letter ‘u’, which is immediately revised, much as we would to when typing on a keyboard.

(36) But a
~~Patriot~~ patriot can be proud of
his ~~**con**~~ country no matter where he
is born or where he lived first.

In addition, revisions were usually interpreted as orthographic revisions in those cases where the deleted elements do not make a complete word, but where the deletions hint at areas of the L2 that are notoriously difficult for learners, such as example (37) which very likely shows difficulties with hyphenation. Having said that, some instances of this kind were interpreted as conceptual changes, i.e. revisions with regard to content, as illustrated in (38). It is plausible that the writer wanted to use the word *unimportant* but changed his/her mind in the middle of writing and chose *unsignificant* instead. (39) illustrates a similar change of mind, this time in the grammatical system of definiteness.

(37) The ground
reason I write text more text ~~**me-**~~
ss messages is because if my friends
have their phone turned off I can't
~~ca~~ reach them.

(38) He feels very ~~unim~~ unsignificant because
he thinks that [...]

(39) [...] they ~~was~~ were
not allowed to go to ~~hos~~ the hospitals for white
Americans.

Table 4 provides a frequency breakdown of the 218 instances of formal-conventional change with regard to the individual categories discussed above. The largest share of revisions are of the grammar type (62; 28%), followed by lexis and orthography as close second (54; 25%) and third (53; 24%). Less frequent, but still prominent are typos with a frequency of 37 tokens (17%).

Tab. 4: A breakdown of formal-conventional revisions with regard to areas of linguistic description.

kind of formal-conventional revision	frequency	
grammar	62	28.4%
lexis	54	24.8%
idiomaticity	7	3.2%
text	5	2.3%
orthography	53	24.3%
typo	37	17.0%
total	218	100.0%

Especially interesting is the comparatively very low frequency of revisions relating to idiomaticity or text, both of which are far less frequent than any of the other four categories, namely 7 (3%) and 5 (2%) tokens, respectively. One explanation might be that these areas of foreign language learning are related to higher levels of proficiency and, therefore, are underrepresented in data that encompass pupils from grade 9 to grade 11 of German secondary schools. With regard to idiomaticity, these figures might mirror the general 'formula-lightness' of learner language as opposed to native language (see, for instance, Granger 1998b or Ellis 2012). However, we cannot neglect the curriculum for English in secondary schools for the federal state of Hesse (relevant for the pupils represented in the MILE) as one contributing factor. While it is true that the curriculum describes as its objectives that pupils should show a high degree of correctness as well as an ability to express themselves idiomatically, it also explicitly states that even for beginner learners particular emphasis should be put on writing with a focus on orthographic and grammatical correctness and lexical appropriateness (Hessisches Kultusministerium 2010: 3–4). After grade 9 pupils are supposed to have an active vocabulary of around 3300 lexical units. Although awareness of topic-related vocabulary, lexical fields and collocations seems to be relevant (Hessisches Kultusministerium 2010: 47), the learning outcomes for the individual grades only once make reference to idiomaticity (in the description for grade 8 we find a mention of phrasal verbs; Hessisches Kultusministerium 2010: 25); the increase of lexical units in terms of numbers seems to be predominant. Orthography still plays an important role in the curriculum for the three final years of secondary school. Again, an increase in lexical competence seems to be primarily measured in terms of the number of lexical units, with a focus on register, genre and cultural awareness (Hessisches Kultusministerium 2010: 10–11). Interestingly, it is only in the three final years that cohesion and coherence become relevant. All this ties in nicely with the results of the analysis presented above.

5.2 Masking problem areas? – Revisions and the final product

It was argued above that revisions provide us with a valuable additional source of data regarding interlanguage. Firstly, any kind of revision can be seen as an indication of problems that a student might have with a particular area of the target language and secondly, the revised end-product might not show any errors although a look at the revisions quickly makes clear that the writer is not as surefooted in the L2 as the final text would lead us to believe. Consider the following examples:

(40) After She splits up with Roger, she begins to
~~**crie**~~ cry ~~a~~ more.
(41) ~~**At the first**~~ The first time he
changes his possitiv attitude into anger about
immigrants is in the first chapter in at the beginning.
(42) people who do
~~x~~ spend ~~a~~ hours ~~**to talk**~~ talking to their
friends on the phone.

As can be seen, this masking concerns different levels of the linguistic system and also different levels of proficiency: whereas the revision in (40) masks a rather basic orthographic problem, (41) and (42) make clear that the writer is not as confident with idiomatic usage as we might assume on the basis of the final products alone. This section will take a look at revisions from the perspective of whether the change improves the text, whether it makes things worse or whether it is rather neutral with regard to correctness.

‘Improvement’ does not necessarily mean that the resulting wording is entirely correct (as in (43)); sometimes it remedies one problem but creates another, less serious problem, as in (44), where the revision corrects the typo but at the same time leads to the incorrect capitalisation of *Involved*.

(43) [...] “Who ~~**ist**~~ is that new guy ~~x~~?” [...]
(44) Getting ~~**ivo**~~ Involved is simple, in the ~~th~~ text Paul gives good examples.

Conversely, erroneous revisions are those which make the resulting sentence ‘worse’. This may mean it renders a correct use of language incorrect (45) or it makes an incorrect sentence even more incorrect (46).

(45) [...] and “legal debate” ~~very~~
sounds official ~~**and**~~ ~~xxx~~ more true than “dreams”.
(46) Reasons for that, mentions the author, are
small electronic devices ~~**have been**~~ $^{\text{are}}$ invent~~x~~$^{\text{ed}}$.

In (45) the writer’s revision deletes the necessary co-ordinator *and*, thus rendering a correct form incorrect. In (46), in contrast, the sentence is incorrect because the relative pronoun after *devices* is missing. In the revision, the writer changes the correct use of the present perfect into the incorrect simple present; not only is the issue not resolved, but another error is added.

Formal-conventional revisions of the ‘neutral’ kind are those that correctly identify an error but replace it by a similarly erroneous version, as in (47) where one incorrect preposition is replaced by another one; *of* or *when* would be correct alternatives, here. In (48), the writer searches for the correct spelling of *through*, but is not successful.

(47) [...] The low angle
shots show the action ~~**by playing**~~ at playing
football and how [...]
(48) And his humour ~~g~~ gets ~~**thou**~~ throg the
hold chapter [...]

Table 5 gives an overview of the frequency of correct, erroneous and neutral revisions of the formal-conventional kind. As Table 5 makes clear, in the vast majority of cases, i.e. over 80%, the revision leads to an improvement of the text. Relatively rarely does the revision make the text worse (30, 13.8%); even more rare

(12, 5.5%) are neutral revisions. We can conclude, then, that an analysis of revisions in authentic learner texts can provide us with a large amount of additional data concerning the interlanguage of the learner. Potential problems that do show in revisions do not show in the final text.

Tab. 5: A breakdown of formal-conventional revisions with regard to success of revision.

kind of formal-conventional revision	frequency	
correct	176	80.7%
erroneous	30	13.8%
neutral	12	5.5%
total	218	100.0%

Tab. 6: The success of formal revisions across revision types with regard to kinds of revisions.

	grammar	lexis	idiomaticity	text	orthography	typo	total
correct	42	43	7	1	47	36	176
erroneous	19	4	0	4	3	0	30
neutral	1	7	0	0	3	1	12
total	62	54	7	5	53	37	218

Finally, it is interesting to see whether the above distribution is more or less similar across the different types of errors identified in the previous section. This information is provided in Table 6. Table 6 shows that some problem areas are more susceptible to revision than others. In particular, typos and orthography problems are usually sorted out in the revising process. The category ‘lexis’, too, mostly shows revisions to be correct or at least neutral; erroneous revision only accounts for 4 out of 54 cases (7.4% of the data), as opposed to 13.8% (see Table 5) across all instances of formal-conventional revisions. The category ‘grammar’ shows a disproportionately large share of unsuccessful revisions (30.7%; 19 out of 62) and a correspondingly low share of correct ones (67.7%; 42 out of 62). These results seem to suggest that, as far as revision goes, grammar is the most demanding area for the L2 writers under analysis. Problems related to spelling, in contrast, are relatively minor in the sense that they are usually remedied during revision.

Interesting results can be seen in the cases of idiomaticity- and text-related revisions. While the omission of revision in both cases coincides with a certain neglect on the part of curriculum designers and, in addition, seems to hint at the challenging nature of these two areas, the opposing tendencies with regard to success of revisions are conspicuous. The high proportion of successful revisions in the first case seems to show some degree of knowledge in the area of idiomaticity whereas the revision failure in the second case shows a general ignorance of text-related matters. Of course, nothing definitive can be said given the low number of tokens in both cases. However, if the tendencies shown in these few instances are borne out in larger data sets and assuming that neither have been taught explicitly, this might indicate that idiomaticity skills are more prone to incidental and implicit learning than textual proficiency. Textual skills seemingly need to be taught through more explicit methods that lead to a cognitivisation of problems and their solutions.

6 Conclusion

The present chapter explored how one approach in writing process analysis, namely the analysis of revisions, can be combined with learner-corpus data. The two usually do not go together since learner corpora mostly provide learner texts as products that provide no information on the process of writing. This is different with the MILE (currently under compilation), which uses a rich text-internal markup system that allows us to partly reconstruct the genesis of the text. The MILE, therefore, is a useful tool to combine the strength of learner-corpus linguistics, namely a large number of informants, with the strength of writing process analysis, namely tapping cognitive processes during writing. The resulting approach, corpus-based revision analysis, can provide us with new insights into the developing interlanguage of language learners.

The analysis showed that conceptual revisions are the most frequent ones, accounting for roughly 46% of the data. Second with about 37% come formal-conventional revisions, i.e. those that are demanded by the L2 language system. These are most interesting with regard to the learners' L2 competence. A breakdown of these revisions into categories reveals an interesting pattern: while revisions relating to grammar, lexis, orthography and typos are fairly frequent, revisions relating to idiomaticity and textual cohesion and coherence are very rare. It was argued that this shows a lack of awareness of these linguistic areas which is in line with the relatively small amount of attention that these topics receive in the curriculum for secondary schools in the federal state of Hesse. As for the success of revisions, the data show that more than three quarters of revisions improve the text, although some problem areas benefit less from revisions than others, e.g. grammar as opposed to lexis and orthography/typos. That is, on the basis of the final product learner may look very proficient, while the amount of formal revisions hints at a potentially large number of problems. It is in this sense that corpus-based revision analysis provides us with a clearer picture of the learners' L2 competence.

Of course, relying on corpus data alone comes with a number of problems. In case of multiple revisions of the same text passage it is not always possible to make claims about the order of revision with absolute certainty, although in many cases there are good indicators for a preferred order. More important and more problematic is the fact that we do not really know why a certain revision was made. The present analysis, therefore, restricts itself to describing whether a particular change, for instance, entails a propositional revision, but it does not claim that this was the writer's motivation for the revision; it is possible, that the writer changes content to avoid language problems. In principle, there are two possible solutions to that problem namely thinking aloud while writing or post-writing interviews. The first is problematic as it makes the writing process less natural and hence the data less authentic. The second solution (suggested for instance by Geisler & Slattery 2007), although a seeming gold standard in writing process research, comes with its own problems, particularly the question of validity of retrospective narratives (Park & Kinginger 2010: 32). Considering the fact that all kinds of data mining in writing process research have their individual problems, the difficulties that come with corpus-based revision analysis seem acceptable, particularly in sight of the vast amount of data that this method can provide.

Where to go from here? A first step, of course, is to extend the existing database. As was described in Section 3, the present analysis took less than five percent of the available data into account. A larger set might provide us with interesting results concerning idiomaticity- and text-related revisions, which were very rare in the present data. In addition, the longitudinal nature of the data in MILE can be exploited. For instance, do we see a change in the relative frequencies of revision categories over time? In particular, do revisions relating to idiomaticity and text become more frequent towards the end of secondary school, indicating an increasing awareness as the writer becomes more proficient in his or her L2? Related to that, it has been argued that formal revisions in text production draw the writer's attention to deficient areas of their L2: "output sets 'noticing' in train, triggering mental processes that lead to modified output. What goes on between the original output and its reprocessed form [...] is part of the process of second language learning" (Swain & Lapkin 1995: 371). Under this assumption, problem areas that lead to frequent revisions in the early stages of learning should show significantly fewer errors in the later stages of learning – a hypothesis that can be tested against the true longitudinal data provided by the MILE. This brings us to the last question: What is the general relation of revisions to errors and vice versa? Do they show a more or less similar distribution among the individual categories, or not? How do they interact: do pupils that revise a lot make lots of errors, showing a fairly high degree of uncertainty regarding the L2, or do revisions reduce the number of errors? Does a low number of revisions coincide with only few errors, indicating an overall high L2 competence? Finally, a closer look at conceptual revisions might prove useful with regard to interlanguage: although it was assumed that these revisions are driven by questions of content, in some cases they may also be the result of an avoidance strategy. If so, they can help us to arrive at a yet clearer picture of the learner's foreign language competence.

On the whole, this chapter has shown that corpus-based revision analysis is a useful tool that combines the strengths of writing process analysis and learner-corpus linguistics, thereby opening up new and promising avenues for future research on second language acquisition.

References

Abdel Latif, M. M. (2008). A state-of-the-art review of the real-time computer-aided study of the writing process. *International Journal of English Studies, 8*(1), 29–50.

Atwell, E., Howarth, P., & Souter, C. (2003). The ISLE Corpus: Italian and German spoken learners' English. *ICAME Journal*, *27*, 5–18.

Barlow, M. (2005). Computer-based analysis of learner language. In R. Ellis & G. P. Barkhuizen (Eds.), *Analysing Learner Language* (pp. 335–357). Oxford, UK: Oxford University Press.

Corder, S. P. (1981). The significance of learner errors. In R. Ellis & G. P. Barkhuizen (Eds.), *Error Analyis and Interlange* (pp. 5–25). Oxford, UK: Oxford University Press.

Ellis, N.C. (2012). Formulaic language and second language acquisition. Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44.

Emig, J. (1971). *The Composing Processes of Twelfth Graders.* Urbana, IL: The National Council of Teachers of English.

Geisler, C., & Slattery, S. (2007). Capturing the activity of digital writing: Using, analyzing, and supplementing video screen capture. In H. A. McKee & D. N. DeVoss (Eds.), *Digital Writing Research. Technologies, Methodologies, and Ethical Issues* (pp. 185–200). Cresskill, NJ: Hampton Press.

Granger, S. (Ed.) (1998a). *Learner English on Computer*. London, UK: Longman.

Granger, S. (1998b). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. In A. Cowie (Ed.), *Phraseology: Theory, Analysis and Applications* (pp. 145–160). Oxford, UK: Oxford University Press.

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Amsterdam, Netherlands: John Benjamins.

Granger, S. (2008). Learner corpora. In A. Lüdeling (Ed.), *Corpus Linguistics. An International Handbook* (pp. 259–275). Berlin, Germany: de Gruyter.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A criticial evaluation. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 13–32). Amsterdam, Netherlands: John Benjamins.

Hasko, V. (2013). Capturing the dynamics of second language development via learner corpus research: A very long engagement. *The Modern Language Journal*, *97*(S1), 1–10.

Hessisches Kultusministerium (2010). *Lehrplan Englisch. Gymnasialer Bildungsgang – Jahrgangsstufen 5G bis 9G und Gymnasiale Oberstufe*. Hessen, Germany: Kultusministerium.

Jäkel, O. (2010). *The Flensburg English Classroom Corpus (FLECC): Sammlung authentischer Unterrichtsgespräche aus dem aktuellen Englischunterricht auf verschiedenen Stufen an Grund-, Haupt-, Real- und Gesamtschulen Norddeutschlands*. Flensburg, Germany: Flensburg University Press.

Kreyer, R. (2015). MILE – The Marburg Corpus of Intermediate Learner English. In M. Callies & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment* (pp. 13–34). Amsterdam, Netherlands: John Benjamins.

Leech, G. (1998). Learner corpora: What they are and what can be done with them. In S. Granger (Ed.), *Learner English on Computer* (pp. xiv-xx). London, UK: Longman.

Lindgren, E. (2005). *Writing and Revising. Didactical and Methodological Implications of Keystroke Logging*. Umeå Universitet, Sweden: Institutionen för Moderna Språk.

New, E. (1999). Computer-aided writing in French as a foreign language. A qualitative and quantitative look at the process of revision. *The Modern Language Journal*, *83*(1), 80–97.

Ortega, L., & Byrnes, H. (Eds.) (2008). *The Longitudinal Study of L2 Capacities.* New York, NY: Routledge.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45.

Park, K., & Kinginger, C. (2010). Writing/thinking in real time. Digital video and corpus query analysis. *Language Learning and Technology*, *14*(3), 31–50.

Raimes, A. (1991). Out of the woods. Emerging traditions in the teaching of writing. *TESOL Quarterly*, *25*(3), 407–430.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10*(3), 209–241.

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages. A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, *15*(3), 201–233.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate. A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391.

Tono, Y. (2012). International Corpus of Crosslinguistic Interlanguage: Project overview and a case study on the acquisition of new verb co-occurrence patterns. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (pp. 27–46). Amsterdam, Netherlands: John Benjamins.

Van Waes, L., Leijten, M., Wengelin, Å., & Lindgren, E. (2012). Logging tools to study digital writing processes. In W. V. Berninger (Ed.), *Past, Present and Future Contributions of Cognitive Writing Research to Cognitive Psychology* (pp. 507–533). New York, NY: Taylor and Francis.

Van Waes, L., & Schellens, P. J. (2003). Writing profiles. The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics*, *35*(6), 829–853.

Voss, R. (1983). Janet Emig's *The Composing Processes of 12th Graders*. A Reassessment. *College Composition and Communication*, *34*(3), 278–283.

Sylvi Rørvik

Marked themes in advanced learner English

Abstract: This chapter investigates the use of marked themes in argumentative texts written by advanced Norwegian learners of English. The learner texts, taken from the Norwegian component of the International Corpus of Learner English, are compared with expert and novice L1 material in English and Norwegian, following the procedure of the Integrated Contrastive Model (Granger 1996). The results show that Norwegian learners underuse marked themes as compared to expert native speakers of English. Further, the learners overuse adverbials as marked themes, and underuse complements. This is partly the result of an underuse of quoted speech in thematic position. The Norwegian learners thus have not mastered the discourse conventions of English when it comes to the use of marked themes, but this is not due to transfer from Norwegian. Developmental factors are shown to be a more important factor influencing the learners.

1 Introduction

This study explores the use of 'marked themes' in argumentative English texts produced by advanced Norwegian learners. In a clause, the theme "is that which locates and orients the clause within its context" (Halliday 2004: 64). As such, it frames the clause and lets the reader know what the starting point of the message is. The theme extends from the beginning of the clause up to and including the first participant, process, or circumstance (Halliday 2004: 79), and the default unmarked option in a declarative clause is that the theme is conflated with the subject. If any other clause element is chosen as the starting point, the clause has a marked theme (Halliday 2004: 73; see further Section 3.3 below). Marked themes thus constitute one aspect of thematic structure, which in Hallidayan systemic-functional grammar is one of several resources for creating texture. The concept of 'texture' is central to the creation of texts. Indeed, the presence of texture is what creates a text from what would otherwise be a random collection of sentences (Halliday & Hasan 1976: 2).

Sylvi Rørvik, Inland Norway University of Applied Sciences, sylvi.rorvik@inn.no

https://doi.org/10.1515/9783110489071-015

The rationale for investigating the use of marked themes in advanced Norwegian-produced learner English is two-fold: firstly, previous studies (see e.g. Hasselgård 1997, 2004, 2005) have suggested that there are different discourse conventions as regards the use of marked themes in English and Norwegian, so it is possible that Norwegian learners of English would be influenced by transfer from their L1 when writing in English. Secondly, several studies have suggested that thematic structure may be a problem area for learners from different L1 backgrounds (cf. e.g. Hawes & Thomas 1997; Green, Christopher, & Mei 2000; Hannay & Martínez Caro 2008; Chen 2010; Hasselgård 2009a, 2009b; Rørvik 2013), so it seems that this is a fruitful area of study in terms of learner English in general. Previous research in this area will be further discussed in Section 2.

The present study adds to the existing literature discussed in Section 2 by comparing the use of marked themes in argumentative texts written by advanced Norwegian learners with a comparable, argumentative L1 English text type, to ensure a fair evaluation of the learner output, and with a comparable, argumentative L1 Norwegian text type, to identify potential transfer-related effects. Further, the Norwegian learner texts are compared with argumentative novice (i.e. student) L1 material in English and Norwegian, to check for any developmental factors that might influence the L2 writers. The procedure follows an adapted version of the Integrated Contrastive Model (Granger 1996, Gilquin 2000/2001), which is presented in Section 2.3.

The study aims to answer the following research questions:

- Do Norwegian learners adapt to L1 English discourse conventions as regards the use of marked themes? If not, which features are different in the learners' output?
- Can any differences between the learner texts and L1 texts be explained by transfer from the learners L1 or by developmental factors?

Section 2 provides an overview of previous research on thematic structure in English and Norwegian, and in novice and learner language, as well as an introduction to central features of the Integrated Contrastive Model. In Section 3 there is an overview of the data and method used, and an outline of the analytical framework employed. Section 4 presents and discusses the findings, and Section 5 provides a summary and conclusion.

2 Previous research on thematic structure

This section has been divided into three subsections. The first of these deals with previous research on thematic structure in English and Norwegian, the second with evidence from studies on thematic structure in novice and learner language, and the final section introduces the Integrated Contrastive Model, which forms the starting point for the design of the present study.

2.1 Thematic structure in English and Norwegian

There are of course numerous studies of thematic structure in English, but this overview has been limited to studies comparing English and Norwegian, since the main aim is to outline features that may cause transfer-related problems for Norwegian learners of English.

In a series of studies based on material from the English–Norwegian Parallel Corpus (ENPC), it has been shown that there are different discourse conventions in English and Norwegian when it comes to thematic structure. Although the languages have similar structural options, there appear to be different preferences with regard to the placement of non-subjects in clause-initial position: Norwegian has a more flexible word order than is the case for English, in the sense that adverbials, objects, and complements are more frequent in initial position in Norwegian clauses (Hasselgård 1997, 1998, 2004, 2005). The relevance of these findings for the present study is that any occurrence of adverbials, objects, or complements in clause-initial position would be considered a marked theme, and, further, that the greater frequency of such constituents in initial position in Norwegian has the effect of making them less marked than in English (Hasselgård 2004: 188–189). Thus, one might hypothesize that Norwegian learners of English, if they were influenced by the conventions of their first language, would overuse marked themes in English. If so, these cases of overuse would not be classified as grammatical mistakes, but the overuse would give the learner texts a non-native flavor. As these contrastive studies are based on fiction (and to a lesser extent non-fiction texts), however, it remains to be seen whether the cross-linguistic differences hold true for the argumentative texts under scrutiny in the present study.

2.2 Thematic structure in novice and learner language

Previous studies have looked at thematic structure in texts written by learners from a number of L1 backgrounds. For instance, in a study of texts written by

Dutch and Spanish learners (material from the International Corpus of Learner English (ICLE)), Hannay and Martínez Caro (2008: 237–238) found that both learner populations underuse marked themes compared to native speakers (NS). It should be noted, however, that the NS material used as the basis for comparison was not of the same text type as that found in ICLE, namely argumentative texts. Hannay and Martínez Caro (2008) compiled an NS corpus consisting of, among other things, in-flight magazine articles, charity mail and academic texts, so it is possible that the underuse of marked themes found in the learner data may be a result of text-type differences. The same reservations may hold for the results reported by Hawes and Thomas (1997: 52–53) for Malaysian students, whose essays contained different proportions of various thematic structures than were found in newspaper texts from *The Times* and *The Sun*, and for Hannay's (2007) study of Dutch ICLE material. What all these studies have in common, however, is the conclusion that learners use similar thematic devices to those employed by native speakers, but that they use them in different ways. Similar findings were reported by Green, Christopher and Mei (2000) for written learner English from Hong Kong, and by Chen (2010) for spoken learner English from China: Chen compared spoken English produced by Chinese learners from the LINDSEI corpus with comparable L1 speech from the LOCNEC corpus, and found that Chinese learners use marked themes more frequently than native speakers (2010: 83).

When it comes to Norwegian learner English, there are two previous studies in particular that are relevant to the present investigation. Hasselgård (2009a) investigated temporal and spatial structuring in texts from the Norwegian part of the ICLE corpus (NICLE), and compared them with texts from LOCNESS, which is the NS component of ICLE, as well as with a self-compiled corpus of L1 Norwegian argumentative essays. The main finding was that Norwegian learners of English transfer thematic patterns from their L1, which leads to an overuse of initial adverbials in their English texts (Hasselgård 2009a: 103). Similar results were found in a more general study of thematic choice in the NICLE material (Hasselgård 2009b), but the author inserts a caution due to the possible incompatibility of the text types investigated. The caveat is that the NICLE texts are compared with material from the ENPC, and this may be the cause of the learners' apparent underuse of fronted objects, which occur mainly in the form of fronted direct speech (Hasselgård 2009b: 124–125), so "[t]here is a particular need for investigating more genres of spoken and written Norwegian, including argumentative prose by professionals as well as students" (Hasselgård 2009b: 136–137). The present investigation seeks to address this issue by including comparable, argumentative writing by both experts and novices in L1 English and L1 Norwegian, in a procedure following an adapted version of the Integrated Contrastive Model (cf. Section 2.3).

2.3 The Integrated Contrastive Model

The Integrated Contrastive Model was introduced by Granger (1996) and elaborated on by Gilquin (2000/2001). It outlines procedures for investigating learner language and trying to account for differences between learner output and NS output. The present investigation employs a slightly adapted version of the original model, so the discussion of the underlying procedural principles will take the adapted version presented in Figure 1 as the starting point.

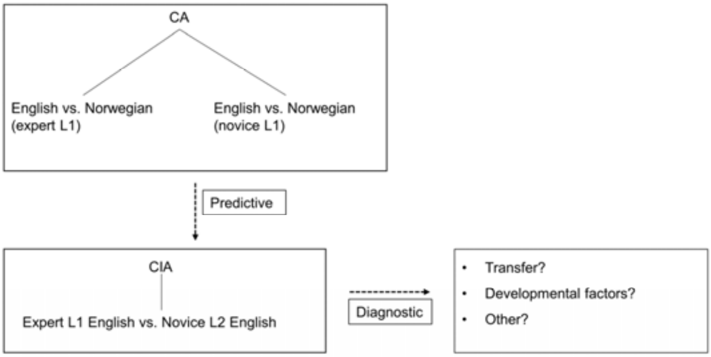

CA
English vs. Norwegian (expert L1)
English vs. Norwegian (novice L1)
Predictive
CIA
Expert L1 English vs. Novice L2 English
Diagnostic
• Transfer?
• Developmental factors?
• Other?

Fig. 1: Adapted version of the Integrated Contrastive Model (Rørvik 2013: 17).

In this model, the evaluation of learner production rests on an initial contrastive analysis of the learners' L1 and the target language, English. The rationale behind this starting point is that one may identify potential problem areas for learners in those areas where their L1 differs from the target language. Note, however, that this is merely a hypothetical prediction, since "L1:L2 mismatches do not always lead to errors, just as L1:L2 identity does not necessarily imply error-free use by learners" (Gilquin 2000/2001: 101). Figure 1 outlines two parallel contrastive analyses of English and Norwegian: one of expert language, and one of novice language. This will be further discussed in Section 3.2.

The next step of the procedure involves a comparison of the learner variety under investigation with the expert L1 material, with the aim of identifying areas of overuse or underuse of linguistic phenomena in the learner output. The adapted version of the model differs from the original model in these first two steps in that the initial contrastive analysis is based only on a comparison of original-language texts in the two languages involved, and in that the only English interlanguage variety explored is produced by Norwegian learners. The final step is diagnostic in nature, and attempts to account for any differences identified between the learners' interlanguage and native-speaker usage. It is of course impossible to investigate and check for all potentially influencing factors,[1] so the figure only lists the two that can be examined directly through a study designed along the lines discussed here, namely transfer and developmental factors. Transfer can be investigated through a comparison of the interlanguage texts with expert texts in the learners' L1 (Gilquin 2000/2001: 101, see also e.g. Hasselgård 2009a, 2009b, and Chen 2010). Developmental factors can be investigated by comparing the interlanguage material with novice writingboth in the target language and in the first language of the learners. In the case of thematic structure and discourse organization, previous studies have shown that this is an area that may cause problems for novice L1 writers as well (cf. e.g. Berry 1995), and studies have shown that "some students apparently [have] developmental problems in argumentative writing not only in L2 but also in L1" (Hirose 2003: 203–204, see also Rørvik & Egan 2013).

3 Data and method

This section provides an overview of the data (3.1) and methods used in the present investigation, together with a detailed outline of the analytical framework employed in the coding of the data (3.2).

3.1 Data

The main focus of the present investigation is the English output produced by advanced Norwegian learners, but one cannot evaluate these texts in a vacuum, so several types of data are necessary in order to answer the research questions outlined in Section 1 (as is also suggested by the description of the Integrated

[1] Other factors potentially influencing learners include, for instance, the effect of teaching, but information regarding this is usually not available from the metadata accompanying corpora, so any conclusions drawn about such matters would be purely speculative in nature.

Contrastive Model in Section 2.3). In total, the study includes five different subcorpora: the L2 English interlanguage material, L1 expert texts in English and Norwegian, and L1 novice texts in English and Norwegian.

The Norwegian learner material has been extracted from the Norwegian component of the International Corpus of Learner English (ICLE; Granger et al. 2009), and will henceforth be referred to as 'NICLE'. This corpus contains 316 argumentative essays (of which 100 have been included in the present study) written by students of English in higher education in response to such prompts as "Crime does not pay" and "Feminism has done more harm to the cause of women than good". As these titles suggest, the students who provided the material for this corpus were not asked to write academic texts, but to give their own personal opinions about an issue. This has implications when it comes to what kind of native-speaker material can be considered as a suitable yardstick, and indeed it has been shown that academic writing is not a good match for the text type found in the ICLE corpus (see e.g. Gilquin & Paquot 2008 and Rørvik 2017, as well as the quotation from Hasselgård 2009b: 136–137 in Section 2.2). Thus, for this study it was decided to compile two subcorpora of expert L1 material in English and Norwegian from the editorials and opinion columns of two online newspapers: *The Guardian*, for the English material, and *Dagbladet*, for the Norwegian material. It should be emphasized that these texts are not news reports or any other text type commonly found in newspapers, which would not be comparable texts to the NICLE material. Instead, they are explicitly expressions of their author's (or authors') opinions, and as such probably the closest one can get to a comparable text type produced by proficient adult writers. In the following discussion, the English expert subcorpus will be referred to as 'ENGNEWS', and the Norwegian expert subcorpus will be referred to as 'NORNEWS'.

While the L1 expert corpora may be considered to represent the target standards to which the writers aspire in their first language and in English, and thus allow for an analysis of transfer-related effects (cf. Sections 2.3 and 3.2), they cannot support any conclusions about potential developmental effects influencing the NICLE writers. For this, comparable novice L1 corpora in English and Norwegian are needed. There is an existing novice NS equivalent to ICLE called LOCNESS (short for the Louvain Corpus of Native English Essays),[2] which contains argumentative essays written by university-age L1 speakers of British and American English, but no such corpus exists for Norwegian novice material. The fifth and final category employed in the present investigation therefore had to be compiled in a more opportunistic way, by downloading argumentative essays written

[2] See http://www.uclouvain.be/en-cecl-locness.html.

in Norwegian by Norwegian high-school students from a website where students can upload their own texts.[3] These L1 Norwegian novice texts will be referred to as ‘NORESSAYS’.

Tab. 1: The number of words in each subcorpus.

Subcorpus	Number of words
NICLE	66,695
ENGNEWS	77,443
NORNEWS	88,010
LOCNESS	86,786
NORESSAYS	102,820
Total	421,754

All five subcorpora described above comprise a total of 100 complete texts, but they vary somewhat in length in terms of the number of words they contain, as shown in Table 1. As is evident from Table 1, the average text length varies between the subcorpora, with the longest texts occurring in NORESSAYS and the shortest in NICLE.

3.2 Method and analytical framework

There is no way of automatically tagging texts for thematic structure, so the texts in each subcorpus were analysed manually by reading through the texts. The definition of theme applied was briefly introduced in Section 1, where it was stated that in a declarative clause, a marked theme involves a non-subject as theme, following Halliday (2004: 79). Thus, in example (1) we find an unmarked theme because the initial constituent is the subject.[4]

(1) I can only hope that readers experienced the emotional journey we took together. (ENGNEWS)

[3] The URL of this website is www.skoleforum.com/stiler/resonnerende.

[4] In all examples the themes have been underlined.

In example (2), on the other hand, we find a marked theme because the sentence starts with an adverbial. In practice, the main options for marked themes are circumstance/adjunct adverbials, and various types of complements, such as objects and predicatives.

(2) In 55 BC, Lucretius speculated that the motion of atoms might be energetic enough to propel them into parallel worlds. (ENGNEWS)

There is one important area in which the present approach differs from Halliday's, and that is in the choice of unit of analysis. This chapter follows Fries (1995: 318) in basing the analysis on 'T-units', i.e. clause complexes consisting of a main clause and all associated hypotactic (dependent/subordinate) clauses. This approach differs from Halliday's practice, where the theme of each ranking clause, whether dependent (hypotactic) or independent is considered (2004: 64–65), and it has two implications for the analysis of theme. The first is that compound sentences will be analysed as two (or more) T-units, i.e. a T-unit has only one main clause, as illustrated in example (3), where the two main clauses have one theme each.

(3) Some may call this argument prejudicial,
but those who do have never spent a morning with Helen Hunt, listening to her expound on her skills. (ENGNEWS)

(4) If they became clearer, the multiverse interpretation might prove unnecessary. (ENGNEWS)

The second implication is that initial dependent clauses will be considered the theme for the T-unit, as illustrated by example (4). Because such initial dependent clauses tend to function as non-subjects, the practice of considering them the theme of their clause complexes results in a greater potential frequency of marked themes. If every ranking clause were considered separately in the thematic analysis, the proportion of thematic subjects would probably increase. This should be borne in mind when comparing the results of the present investigation with those from previous research, whose units of analysis vary greatly. The major advantage of using the T-unit in the present study, however, is that it ensures comparability with previous contrastive studies of Norwegian and English, and of Norwegian learner English (cf. e.g. Hasselgård 1997, 1998, 2004, 2005 and Rørvik 2013). Table 2 contains an overview of the number of T-units in each subcorpus.

Tab. 2: The number of T-units and marked themes in each subcorpus.

Subcorpus	Number of T-units	Number of marked themes
NICLE	4,313	835
ENGNEWS	4,167	959
NORNEWS	5,674	1,210
LOCNESS	5,371	1,236
NORESSAYS	7,198	1,653
Total	26,723	5,893

All marked themes identified in stage one of the analysis were further categorized according to their syntactic function as either adverbial or complement (for non-adverbials, i.e. objects and predicatives). The form of the marked themes was also coded, which naturally entailed letting the categories emerge from the data. The following structures were found: prepositional phrase, dependent clause, noun phrase, adverb phrase, adjective phrase, and quoted speech.

The further analysis followed the principles of the Integrated Contrastive Model, as outlined in Section 2.3. The NICLE texts were compared with the L1 expert English texts, and any differences between the learners and the L1 expert writers were then scrutinized further to attempt to identify factors influencing the learners. If the differences between the NICLE texts and the ENGNEWS texts were similar to differences between the ENGNEWS and the NORNEWS texts, the possibility of transfer could not be excluded. If, on the other hand, the other novice corpora differed from the expert texts in similar ways, but no cross-linguistic difference was found, this was taken as an indication that developmental factors might be at play. All frequencies were subjected to statistical testing in the form of a one-way ANOVA with a Tukey post-hoc test. The results of the analyses are presented in Section 4.

4 Findings

The discussion of the findings has been divided into three subsections. The first presents the results for the overall frequencies of marked themes, the second focuses on the syntactic function of the marked themes, and the third on the structural realization of the marked themes.

4.1 The frequency of marked themes

The proportion in percent of marked themes in each subcorpus is shown in Figure 2. The boxplots in Figure 2 indicate a fairly similar degree of corpus-internal variation in each subcorpus, i.e. the writers within each corpus differ from each other to a similar extent.

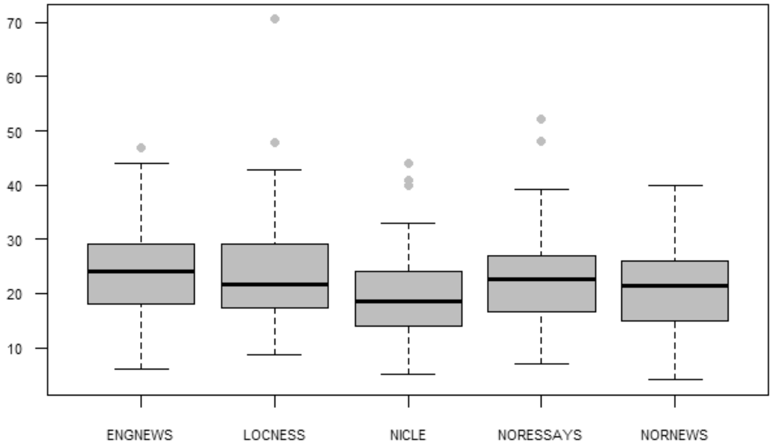


Fig. 2: The proportion of marked themes in each subcorpus (percentage of total themes).

The median (represented by the black line in each plot) is lower in NICLE than in the English L1 expert corpus, and the statistical testing shows that the NICLE writers use significantly fewer marked themes than the ENGNEWS writers (p=0.0131779). This is a similar result to that reported by Hannay and Martínez Caro (2008) for Dutch and Spanish ICLE writers, but it seems that the Norwegian ICLE writers are different from Chinese learners, given Chen's (2010) results discussed in Section 2.2. Chen found that Chinese learners use marked themes more frequently than native speakers, but she used spoken material for her study, so the difference may be due to the mode of communication investigated.

The issue of what influences the learners to underuse marked themes is complicated. It was suggested in Section 1 and in some of the previous research discussed in Section 2 that there are different discourse conventions in Norwegian and English which might influence the learners, in the sense that they would transfer patterns from Norwegian to their English output. However, although the medians for ENGNEWS and NORNEWS are slightly different, there is no significant difference between the two corpora (p=0.3100808). We can therefore exclude the possibility of transfer from the learners' L1 as the cause of the underuse of marked themes. The design of the study also allows for the investigation of the effect of developmental factors, but again there is no clear evidence for this, since the novice L1 corpora are not significantly different from the L1 expert corpora: the comparison of LOCNESS with ENGNEWS yielded a p-value of 0.9995715, and the comparison of NORESSAYS with NORNEWS resulted in a p-value of 0.7723513. Thus, the data cannot support any conclusions regarding the cause of the underuse of marked themes in NICLE. It may be possible, however, to pinpoint the specific areas where the underuse occurs, so we now turn to the syntactic function of the marked themes.

4.2 The syntactic function of the marked themes

There are two possible syntactic functions of the marked themes: adverbial and complement. We will start with the most frequent function, namely adverbials, and Figure 3 contains the percentage of all marked themes that function as adverbials in each corpus.

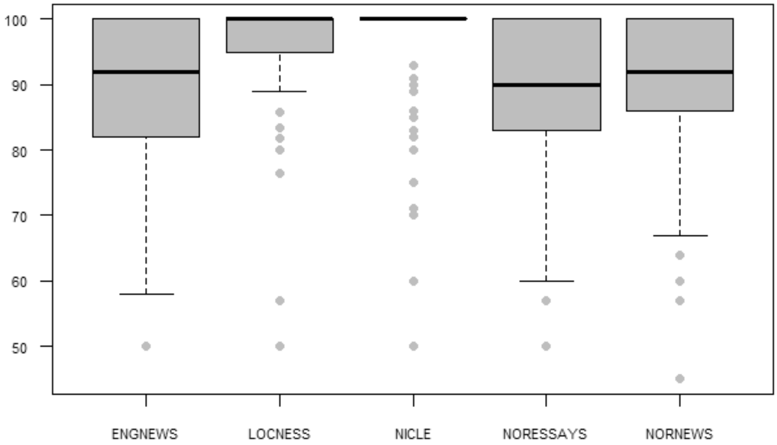


Fig. 3: The proportion of marked themes functioning as adverbials in each subcorpus (percentage of total marked themes).

It is clear from Figure 3 that the NICLE writers, apart from a relatively low number of outliers, almost exclusively use adverbials as marked themes. The contrast to the proportion of adverbials in ENGNEWS is obvious, and statistically significant (p=0.0004331). This overuse of adverbials by the Norwegian learners ties in with the findings from previous studies (Hasselgård 2009a: 95, 2009b: 124), but the statistical analysis shows that this overuse in NICLE cannot be attributed to transfer, since there is no significant difference between the two L1 expert corpora (p=0.9983360). This may be taken as an indication that the use of a similar text type as the baseline is important, since Hasselgård (2009a) concluded that the NICLE writers' overuse was due to transfer from Norwegian, on the basis of a contrastive analysis of ENPC material. As regards developmental factors, the Norwegian L1 novices are not significantly different from the Norwegian expert writers (p=0.6567868), and indeed they clearly use a smaller proportion of adverbials than the NICLE writers do. However, a comparison of LOCNESS with ENGNEWS shows that the L1 English novices also overuse adverbials in thematic position (p=0.0001823), so it is possible that this overuse is a feature of novice English writing, regardless of whether the writers produce texts in their first language or in a foreign language.

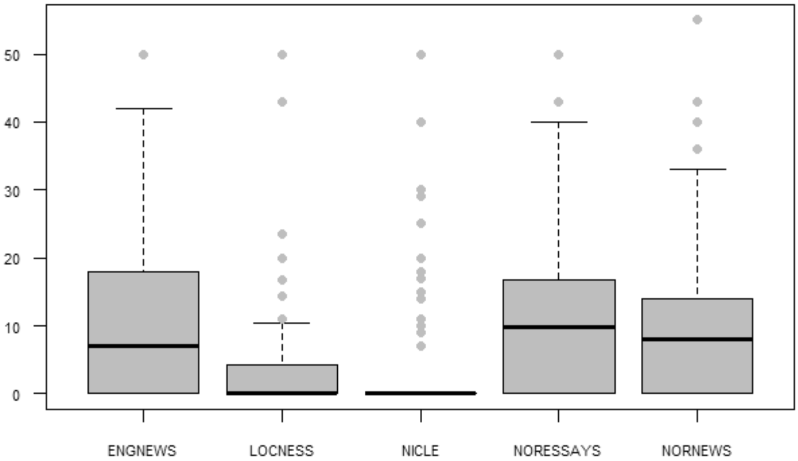


Fig. 4: The proportion of marked themes functioning as complements in each subcorpus (percentage of total marked themes).

The proportions of thematic complements, illustrated in Figure 4, must necessarily be a mirror image of the results presented in Figure 3. Given the results presented in Figure 3, it is not surprising that Figure 4 shows underuse of complements by the NICLE writers compared to the ENGNEWS writers (p=0.0012155).

Again, there is no evidence of the cross-linguistic difference reported by Hasselgård (1997, 1998, 2004, 2005), since the proportion of complements is not significantly higher in the Norwegian expert material than in the English expert material (p=1.0000000), which yet again underlines the importance of the choice of text type. It does seem that developmental factors play a role with regard to the proportion of complements as well, however. The LOCNESS writers use significantly fewer thematic complements than the English expert writers (p=0.0003225), so novice writers of English share the tendency to underuse complements as themes.

Thus far we have seen that the English interlanguage of Norwegian learners is characterized by an underuse of marked themes, and that, out of the marked themes they do use, too many are adverbials and not enough are complements. Section 4.3 attempts to account for these differences by investigating the structures used as marked themes.

4.3 The form of the marked themes

As mentioned in Section 3.2, there were six structures that functioned as marked themes in the material: prepositional phrase, dependent clause, noun phrase, adverb phrase, adjective phrase, and quoted speech. However, the frequency of adjective phrases was so low that they will be left out of the present discussion.

Figure 5 shows the proportion of marked themes realized by prepositional phrases. Prepositional phrases are the most frequent realizations of marked themes in all five subcorpora, and although there is more corpus-internal variation in NICLE than in the other corpora, the general tendency is that the NICLE writers adapt well to the English discourse conventions with respect to the use of prepositional phrases as themes. Indeed, there is no significant difference between the NICLE texts and the English expert texts (p=0.8533359). It is possible that the NICLE writers are helped by what may be termed positive transfer in this case, since the L1 expert corpora are not significantly different from each other (p=0.1722344).

The use of dependent clauses as theme is another area where the Norwegian learners successfully adopt English usage patterns. As illustrated in Figure 6 there is slightly more corpus-internal variation in NICLE, but once again the difference between the Norwegian learners and the L1 English experts is not significant (p=0.0739641). In fact, when it comes to dependent clauses in thematic position, the NICLE writers have successfully navigated an area where there is a cross-linguistic difference between English and Norwegian, because there are significantly fewer such themes in NORNEWS than in ENGNEWS (p=0.0006333).

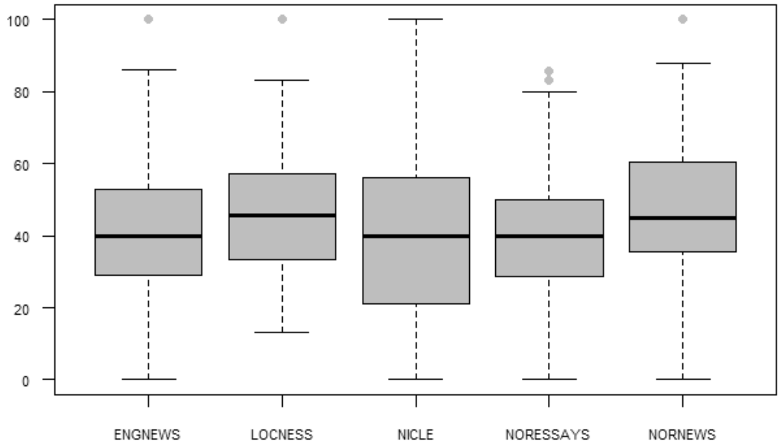


Fig. 5: The proportion of marked themes realized by prepositional phrases in each subcorpus (percentage of total marked themes).

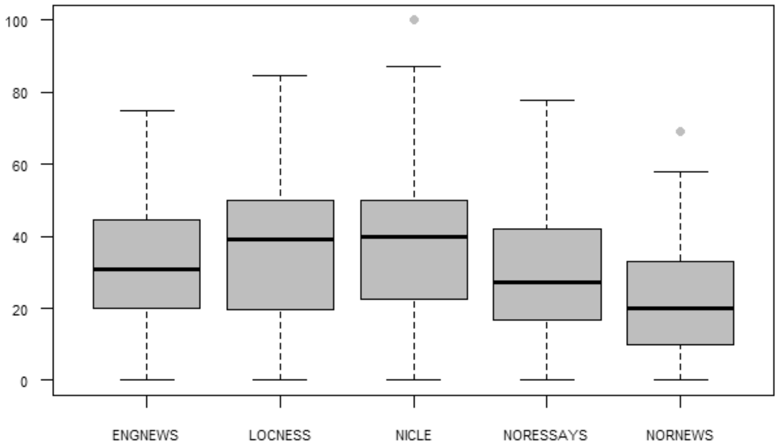


Fig. 6: The proportion of marked themes realized by dependent clauses in each subcorpus (percentage of total marked themes).

As regards noun phrases functioning as marked themes, the situation is similar to that discussed in relation to prepositional phrases, although noun phrases are much less frequent, as can be seen in Figure 7. The NICLE writers use thematic noun phrases to the same extent as L1 expert writers of English (p=0.7248247), and are possibly helped by the lack of a significant difference between Norwegian and English (p=0.8303302), despite the fact that there is a tendency for Norwegian writers to use noun phrases in this way to a slightly greater extent than English writers do.

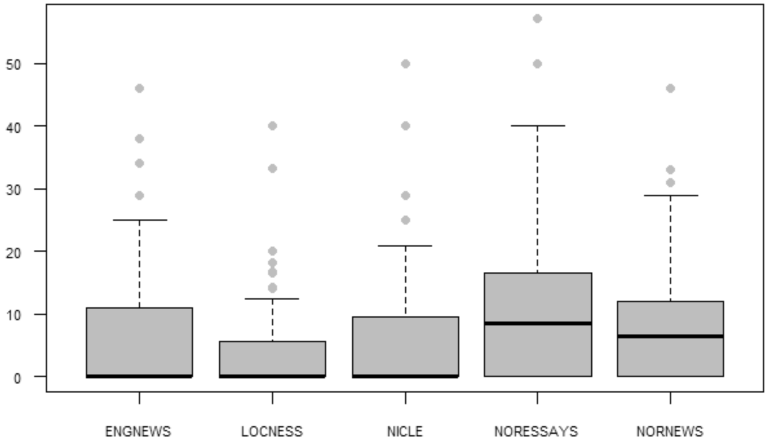


Fig. 7: The proportion of marked themes realized by noun phrases in each subcorpus (percentage of total marked themes).

Adverb phrases account for a larger proportion of marked themes than noun phrases, and Figure 8 shows that there is also more corpus-internal variation in the use of adverb phrases than was the case for noun phrases. There do seem to be slightly higher proportions of adverbs in the NICLE data, but this is yet another thematic device which the Norwegian learners use with a frequency that is not significantly different from the L1 English expert writers (p=0.2461857), despite the fact that there is evidence to suggest a cross-linguistic difference, based on the comparison of ENGNEWS and NORNEWS (p=0.0015179).

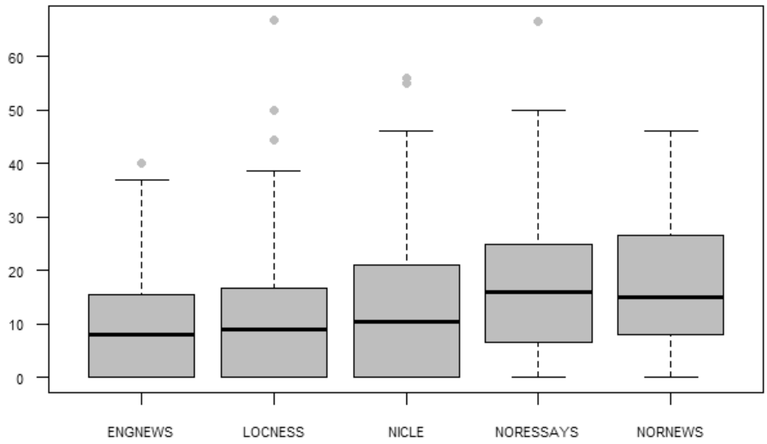


Fig. 8: The proportion of marked themes realized by adverb phrases in each subcorpus (percentage of total marked themes).

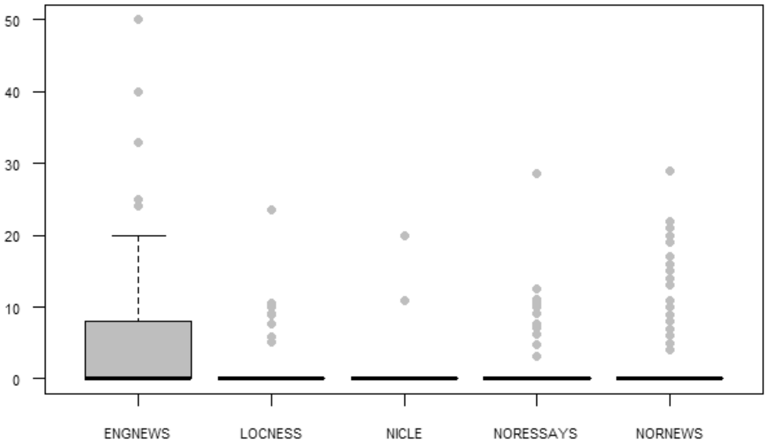


Fig. 9: The proportion of marked themes realized by quoted speech in each subcorpus (percentage of total marked themes).

Finally, we will look at the use of quoted speech as marked themes. The distribution is shown in Figure 9. It is immediately clear from Figure 9 that a major cause of the underuse of complements in NICLE, as discussed in Section 4.2, is an underuse of marked themes realized by quoted speech, and this underuse may also go some way towards explaining the general underuse of marked themes in NICLE. ENGNEWS contains significantly more than NICLE of this type of marked theme (p=0.0000015), so this feature warrants further attention. Examples (5) through (9) illustrate the phenomenon.

(5) "Markets need morals," said Gordon Brown in his Labour conference speech yesterday. (ENGNEWS)
(6) "I have never had the ambition of power," he said at the time. (ENGNEWS)
(7) "This is ridiculous, they'll never press charges," lawyers who attended to the arrested said on the day. (ENGNEWS)
(8) "We are having to balance energy savings with customer concern," says Walker-Palin. (ENGNEWS)
(9) "Give me a Kalashnikov," said the Talib. (ENGNEWS)

In all of these examples, the writer has chosen to place the quoted speech first in the sentence. In some cases, such as example (7), for instance, this choice may be related to the principle of end weight. If the writer of example (7) had chosen to start with the subject ("lawyers who attended to the arrested"), the resulting sentence would have been hard to read. In the case of the other examples provided here, the choice would not have had such dramatic results for the balance of the sentences, but possibly the writers felt that starting with the quote provided the opportunity for a more dynamic reproduction of events. There is also the question of thematic progression to consider: it is possible that this order of elements was a better fit with the preceding co-text, in terms of the information principle and the topical progression of the texts.

There is a parallel here to the results reported by Hasselgård (2009b), who also found underuse of fronted direct speech by NICLE writers, and attributed this to text-type differences, since the L1 material used for comparison had been taken from the ENPC. It is perhaps tempting to attribute the similar underuse found in the present investigation to the same cause, and conclude that the opinion pieces published in newspapers would more often refer to events in the outside world than the NICLE writers do. This becomes an even more plausible explanation when one takes into account the fact that the expert L1 writers do not differ significantly from each other in the use of quoted speech (p=0.1196657), and that the LOCNESS writers are also underusing this feature compared to the English expert writers (p=0.0000177). However, the novice L1 Norwegian writers do not differ significantly from the expert Norwegian writers when it comes to the proportion of marked themes realized by quoted speech (p=0.3718111), and therefore it cannot simply be a question of argumentative student essays being a different text type to that found in newspaper opinion pieces, although this may have some effect on the results. It would seem, however, that yet again the novice writers of English, both L1 and L2, share a feature, and, in fact, that it is this characteristic that underlies their similarities in the underuse of initial complements and therefore overuse of initial adverbials discussed in Section 4.2.

5 Discussion

This section discusses the main findings of the present paper in the light of previous research. As regards the frequency of marked themes, previous research has identified cross-linguistic differences between Norwegian and English in fiction texts (Hasselgård 1997, 1998, 2004, 2005) in the potential for marked themes, and suggested that Norwegian discourse conventions favour marked themes to a greater degree than English conventions. If the Norwegian learners were influenced by their L1, we might then hypothesize that they would overuse marked themes in English. However, as shown in Section 4, the present investigation found no cross-linguistic difference, but underuse of marked themes among the Norwegian learners. This underuse mirrors the tendencies identified by Hannay and Martínez Caro (2008) in Dutch and Spanish learner material, but is the opposite of what Chen (2010) found in the spoken English of Chinese learners.

When it comes to the syntactic function of marked themes, the present investigation found overuse of adverbials and underuse of complements among Norwegian advanced learners as compared to expert L1 writers of English. These results are similar to the findings presented by Hasselgård, who identified overuse of initial adverbials among Norwegian learners (2009a) and underuse of fronted objects (2009b).

The final feature investigated in the present study was the form of marked themes. This seemed to be an area where the Norwegian learners mostly succeeded in following English discourse conventions, since they did not differ significantly from NS expert writers when it came to marked themes realized by prepositional phrases, dependent clauses, noun phrases, and adverb phrases. However, marked themes realized by quoted speech account for a lower proportion of marked themes in NICLE than in L1 expert English, despite the lack of a contrastive difference. This is similar to Hasselgård's (2009b) results, which also indicated an underuse of fronted direct speech among Norwegian learners, although the learner texts were compared with material from the ENPC in that instance.

6 Conclusion

The aim of the present study was to investigate the use of marked themes in argumentative texts by Norwegian advanced learners of English, and this was operationalized in two research questions. The first of these asked whether Norwegian learners use marked themes in the same way that native speakers of English do. As shown in Section 4, there are a number of differences between the learner output and the English expert material, so it seems clear that the Norwegian learners have not mastered English discourse conventions regarding the use of marked themes. The learners underuse marked themes overall, and in terms of the syntactic functions of the marked themes they overuse marked themes functioning as adverbials and underuse those functioning as complements. A major factor in the underuse of complements is the underuse of quoted speech in thematic position.

The second research question focused on identifying factors influencing the learners in those cases where their thematic choices differ from those made by native speakers. The design of the study means that two such factors can be directly studied, namely transfer from the learners' L1 and developmental factors. The overall underuse of marked themes could not be explained by reference to either of these factors, and indeed transfer from Norwegian seems to be largely irrelevant as an explanation for the features where the learner texts differ from the native-speaker expert texts, which is perhaps surprising given the results of previous studies discussed in Section 2. Developmental factors, however, seem to account for the overuse of marked themes functioning as adverbials, and the underuse of complements and themes realized by quoted speech. This conclusion is based on the fact that the L1 English novice writers in LOCNESS differ from the English expert texts in similar ways to the Norwegian L2 writers. In this sense, the results of the present study corroborate the conclusion drawn by Hirose (2003: 182): "We cannot discuss student L2 organizational patterns without taking into consideration student L1 and L2 writing background in terms of writing conventions, instruction, and experience, as well as L2 proficiency level."

References

Berry, M. (1995). Thematic options and success in writing. In M. Ghadessy (Ed.), *Thematic Development in English Texts* (pp. 55–84). London, UK: Pinter.

Chen, X. (2010). *Discourse-Grammatical Features in L2 Speech: A Corpus-Based Contrastive Study of Chinese Advanced Learners and Natives Speakers of English.* Doctoral dissertation, The City University of Hong Kong, Hong Kong.

Fries, P. H. (1995). Themes, development and texts. In R. Hasan & P. Fries (Eds.), *On Subject and Theme* (pp. 317–359). Amsterdam, Netherlands: John Benjamins.

Gilquin, G. (2000/2001). The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast*, *3*(1), 95–123.

Gilquin, G., & Paquot, M. (2008). Too chatty. Learner academic writing and register variation. *English Text Construction*, *1*(1), 41–61.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies* (pp. 37–52). Lund, Sweden: Lund University Press.

Granger, S, Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English v2*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Green, C. F., Christopher, E. R., & Mei, J. L. K. (2000). The incidence and effects on coherence of marked themes in interlanguage texts: A corpus-based enquiry. *English for Specific Purposes*, *19*(2), 99–113.

Halliday, M. A. K. (2004). *An Introduction to Functional Grammar* (3rd edition, revised by C. M. I. M. Matthiessen). London, UK: Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London, UK: Longman.

Hannay, M. (2007). Patterns of multiple theme and their role in developing English writing skills. In C. Butler (Ed.), *Functional Perspectives on Grammar and Discourse: In Honour of Angela Downing* (pp. 257–578). Amsterdam, Netherlands: John Benjamins.

Hannay, M., & Martínez Caro, E. (2008). Thematic choice in the written English of advanced Spanish and Dutch learners. In G. Gilquin, S. Papp, & M. Belén Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research* (pp. 227–253). Amsterdam, Netherlands: Rodopi.

Hasselgård, H. (1997). Sentence openings in English and Norwegian. In M. Ljung (Ed.), *Corpus-based Studies in English. Papers from the 17th International Conference on English Language Research on Computerized Corpora* (pp. 3–20). Amsterdam, Netherlands: Rodopi.

Hasselgård, H. (1998). Thematic structure in translation between English and Norwegian. In S. Johansson & S. Oksefjell (Eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies* (pp. 145–167). Amsterdam, Netherlands: Rodopi.

Hasselgård, H. (2004). Thematic choice in English and Norwegian. *Functions of Language*, *11*(2), 187–212.

Hasselgård, H. (2005). Theme in Norwegian. In K. L. Berge & E. Maagerø (Eds.), *Semiotics from the North. Nordic Approaches to Systemic Functional Linguistics* (pp. 35–47). Oslo, Norway: Novus Press.

Hasselgård, H. (2009a). Temporal and spatial structuring in English and Norwegian student essays. In R. Bowen, M. Mobärg, & S. Ohlander (Eds.), *Corpora and Discourse – and Stuff. Papers in Honour of Karin Aijmer* (pp. 93–104). Gothenburg, Sweden: Acta Universitatis Gothoburgensis.

Hasselgård, H. (2009b). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 121–139). Amsterdam, Netherlands: Benjamins.

Hawes, T., & Thomas, S. (1997). Problems of thematisation in student writing. *RELC Journal, 28*(2), 35–55.

Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of Second Language Writing*, *12*(2), 181–209.

Rørvik, S. (2013). *Texture in Learner Language*. Doctoral dissertation, University of Oslo, Oslo, Norway.

Rørvik, S. (2017). Norwegian L2 writers' connector use: A great degree of lexical variation, or overuse of high-frequency items? In P. de Haan, S. van Vuuren, & R.de Vries (Eds.) *Language, Learners and Levels: Progression and Variation* (pp. 69–91). Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Rørvik, S., & Egan, T. (2013). Connectors in the argumentative writing of Norwegian novice writers. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead* (pp. 401–410). Louvain-la-neuve, Belgium: Presses universitaires de Louvain.

Susan Nacey and Anne-Line Graedler

Phrasal verbs in the spoken and written modes of Norwegian L2 learner English

Abstract: This chapter explores the use of English phrasal verbs (PVs) by Norwegian L2 learners by investigating data from the Norwegian part of the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI). A total of 1,489 PVs were first identified and analyzed for possible contrast between spoken and written modes. While the findings reveal some differences between both individual L2 learners and also between the spoken and written corpora, the general perception that the use of English PVs is highly problematic for language learners is not supported. The roles of metaphoricity and L1 transfer in relation to divergent PVs were also investigated. Findings suggest that metaphor may provide more help than hindrance to L2 language learners, but we find that more reliable investigation requires larger datasets of L2 learner language than are currently available.

1 Introduction

Language learners' mastering of English phrasal verbs is generally acknowledged as difficult, "one of the most problematic areas for learners of English" (Jenkins 2009: 51). Possible pitfalls are numerous, since phrasal verbs are subject to both syntactic restrictions and semantic challenges, as well as potential first language (L1) transfer. At the same time, phrasal verbs are viewed as important, and even indispensable, to second or foreign language (L2) proficiency in English – a "can't miss topic" (Gilquin 2015: 59), and subject to a fair amount of research.

In the present study, the term phrasal verb (PV) includes both intransitive and transitive [verb + particle] combinations, e.g. *break up* and *fall for something*, as well as verbs combined with a particle and a preposition, e.g. *take care of* (a more detailed explanation in Section 2.1 below). By investigating the PV production in written and spoken language of Norwegian advanced learners of English, the present chapter adds empirical corpus-based evidence concerning the real

Susan Nacey, Inland Norway University of Applied Sciences, susan.nacey@inn.no
Anne-Line Graedler, Inland Norway University of Applied Sciences, anneline.graedler@inn.no

https://doi.org/10.1515/9783110489071-016

magnitude of the challenge that PV use presents, thus comparing PV use across modes and for an additional group of English language learners.

In Section 2 some previous studies are presented as a background for the specific research questions for the present study, as well as a more detailed definition of PVs. Section 3 presents the corpus material used in the investigation and our methodological approach: all identified PVs in the spoken and written data were categorized for their degree of metaphoricity and conventionality. Novel PVs were further explored to assess possible L1 influence as a motivating factor in their production. Section 4 continues with a general overview of PV use, followed by subsections explaining our findings with respect to each of our research questions. Section 5 closes with concluding remarks.

2 Background on phrasal verb use in L2 English

Much of the previous research on PV use in L2 English has focused on learner avoidance of PVs rather than what learners actually do produce (e.g. Hulstijn & Marchena 1989; Laufer & Eliasson 1993; Liao & Fukuya 2004). Some notable exceptions include corpus-based studies, such as Waibel (2007) of written German and Italian L2 English, Mondor (2008) of written Swedish L2 English, Kamarudin (2013) of spoken and written Malaysian L2 English, and Gilquin (2015) of spoken and written French L2 English, all of which compared learners' use of verb-particle combinations with the PV use of native speakers of English.

Waibel (2007) studied two groups of advanced learners with different L1s, German and Italian; she found that in comparison with native speakers of English, Italian students underused PVs in the corpus data, whereas German students overused them. A logical explanation is that German students often use English verbs with Germanic origin, which is the kind of verbs English PVs are generally based on, whereas the Italian students tend to use more verbs with Latin origin (Waibel 2007: 159–160). Similar results were found in Mondor's (2008) study where overuse was linked to the learners' Germanic L1, Swedish, and Gilquin's (2015) study of French learners who used only about half as many PVs as native speakers.

Learners' misuse or deviation from the English standard of PVs, hereafter referred to as divergent PV usage, is related to different levels, such as the lexical choice of specific particles and verbs and their combination, the syntactic structure of the PVs, and the contextual use of PVs, both in relation to collocates and the general style of the text. Hence, several potential challenges exist for L2 learners. First, PVs are subject to syntactic restrictions that may not be readily apparent to learners (cf. e.g. Gilquin 2015). Second, semantics poses a challenge, as PVs are highly polysemous with both literal and (often several) figurative or metaphorical meanings (Kamarudin 2013: 222; cf. also Waibel 2007; Mondor 2008). Third, negative transfer from the learners' L1 may also play a role, especially when it comes to other languages that have verb-particle constructions similar to PVs.

Adding to these challenges are stylistic considerations, since PVs are sometimes relatively informal, which may affect variations between learners' written and spoken texts as well as between different text genres (Waibel 2007: 160; Gilquin 2015). Furthermore, pedagogical practices can also influence learners' use of PVs, since they are often presented to learners either in lists arranged by verb, and/or accompanied by their single-word "equivalents", leading to the mistaken view that all English PVs have complete Latinate synonyms. Even definitions proffered by so-called experts vary, arguably adding to learner confusion (Cowie 1993: 38–39; Waibel 2007: 21–32).

Taken together, these studies indicate that there may be links between L2 PV use and a number of factors, including whether there are PVs in the learners' L1, the opacity of the PV and the degree of its idiomaticity, the PV frequency and register, the proficiency of the learners, and the mode of production. Hence, the present study focuses on the production of PVs in L2 English by advanced Norwegian learners, both by comparing L2 PVs in learner corpus data with L1 standard PV usage, and by attempting to discover some of the reasons for learners' divergent use of English PVs. Four related research questions are addressed:

1. How often do learners produce divergent PVs?
2. Are there contrasting patterns of PV usage across the spoken and written modes?
3. Is there a link between divergent PV usage and metaphoricity?
4. Is there a link between divergent PV usage and L1 transfer?

2.1 What exactly are phrasal verbs?

In the present investigation, PVs are defined as both syntactic and semantic units, following Quirk et al. (1985: 1150–1165). In syntactic terms, a PV consists of a verb followed (immediately or not) by a particle, with no possibility of inserting an adverb between the two elements. This construction may be intransitive (e.g. *go back*) or transitive; if transitive, the particle may typically precede or follow the direct object (e.g. *point X out/point out X*). As mentioned in the introduction above, our working definition of phrasal verbs also includes the [verb + particle

+ preposition] construction, so-called phrasal-prepositional verbs, as in *get away with*. In the active voice, phrasal-prepositional verbs are always transitive, as the preposition triggers the requirement for a direct object, such as *a good story* in (1). In the passive voice, the object is transformed into the subject – e.g. *this program* in (2) – and the preposition is stranded.

(1) I could really write anything if I . just **come up with** *a good story* (NO025)[1]
(2) The dissapointing turn was that *this program* was **given up on**. (UO2041)

In semantic terms, the verb and particle of a PV form a single unit of meaning. There is frequently, though not always, one or more single-word alternatives for PVs (e.g. *leave out* = *omit*). The meaning of the particle somehow modifies that of the verb, even though the overall meaning of the PV may not necessarily be semantically transparent. Such opacity is more acute when the choice of particle is figuratively motivated (e.g. *give in*, *figure out*) than when the motivation is literal (e.g. *lie down*). It is this idiomatic nature of PVs that is especially said to pose problems for English language learners (cf. Section 1), as the motivation for a particular verb-particle combination may seem unclear or even completely random.

Semantic considerations distinguish PVs from prepositional verbs (e.g. *cope with*), where the preposition is determined in relation to the object rather than the verb. However, syntactic criteria distinguishing the constructions are often needed as a supplement; for example, whether the phrase can be fronted, or whether an adverb may be inserted between verb and particle. Semantic considerations also distinguish PVs from so-called "free combinations", where the meanings of the verb and particle are independent of each other (e.g. *walk past*). However, the borderline between free combinations and PVs is fuzzy, something especially noticeable in the cases where the particle is locative rather than idiomatic (e.g. *sit down*, where *down* indicates the actual direction of sitting). Quirk et al. (1985: 1152) argue that the possibility of substitution of another verb in the construction indicates whether verb and particle are separable (i.e. *walk* in *walk past* could be replaced by other verbs of motion).

[1] Examples tagged as 'NO' followed by a number have been extracted from the LINDSEI-NO transcriptions (cf. Section 3.2), and include markers of hesitation and/or disfluency. All others have been taken from NICLE, and are reproduced exactly as they appear in the corpus. Relevant PVs are marked with bold type.

Norwegian, the L1 of the informants in the present study,[2] is genetically related to English, both being Germanic languages. As Askedal (1994: 262–263) explains, Norwegian is similar to English, in that it has various types of composite verbal constructions, including both phrasal and prepositional verbs. As with English, a pronominal PV object always precedes the particle; otherwise the object may either precede or follow the particle subject to some restrictions such as length of object. Further, Norwegian particles may either be homonymous with prepositions, as in the phrase *legge på prisen* (lit.: 'lay on price-the', colloq.: 'increase the price'), or with adverbs, as in *legge fram X* (lit.: 'lay forth X', colloq.: 'present X'). Phrasal-prepositional verbal constructions also exist in Norwegian: *gå med på X* (lit.: 'go with on X', colloq.: 'comply with X'). Norwegian PVs may also express literal or figurative meanings. The greatest difference between Norwegian and English with respect to PVs is that Norwegian sometimes allows a prefixal formation; such constructions may be the semantic equivalent of PVs with adverb-like particles, e.g. *framlegge* (lit.: 'forth-lay', colloq.: 'present').

3 Methodology

The following section includes information about the corpus data under investigation, as well as a description of how the PVs were identified and extracted from the corpora. The chapter continues by presenting the categorization process of the data both for metaphoricity and for divergent usage.

3.1 Corpus data: LINDSEI-NO and NICLE

The data analyzed in this investigation comes from two L2 English learner corpora, one of spoken language and one of written language. The informants were all advanced learners of English whose L1 was Norwegian. The spoken data consists of all PVs uttered by fifty students in the Norwegian subcorpus of the Louvain International Database of Spoken English Interlanguage, LINDSEI (see Gilquin, De Cock & Granger 2010). These PVs were produced in the context of informal interviews with the students, and their contextual meanings were determined through reference to their respective co-text. The Norwegian subcorpus,

[2] Neither of the corpora used in the present study (NICLE and LINDSEI-NO) contains any metadata concerning the informants' local dialect or preferred written standard form of Norwegian (*Nynorsk* or *Bokmål*).

LINDSEI-NO, will form part of the planned second, expanded version of LINDSEI and is currently only available at Inland Norway University of Applied Sciences, where it was compiled. This subcorpus comprises approximately 13 hours of conversation, amounting to roughly 83,000 words of learner-produced text (excluding backchannels and fillers), which was transcribed and then searched for PVs. The written data consists of all PVs produced in argumentative essays in the Norwegian subcorpus of the International Corpus of Learner English (ICLE), hereafter referred to as NICLE. Slightly more than 211,000 words of text from 317 essays were searched for PVs; this is the entire NICLE corpus, together with one additional essay given us by the compiler of the subcorpus. NICLE was collected around the year 2000, while LINDSEI-NO was collected between 2010 and 2012. This investigation does therefore not explore spoken and written texts by only a *single* group of students – that is, we are not comparing the spoken and written PV usage of the same people. The two groups are however comparable in that all the students were enrolled on a 60-credit college course in English language, literature and culture at a Norwegian institution of higher education.

3.2 Data extraction and analysis

Phrasal verbs are notoriously productive in English, and hence it is not possible to list all potential verbs (cf. Waibel 2007: 65). There is, however, a limited number of potential PV particles, advantageous when attempting to identify PVs in a semi-automated manner. In the present project, our first step was to submit the NICLE texts to part-of-speech (PoS) tagging through the Constituent Likelihood Automatic Word tagging Systems (CLAWS),[3] using the C7 tagset. All sentences containing lexical units tagged in this way as 'RP' (meaning 'prep. adverb, particle') were transferred into our database.[4] Each instance was then manually combed through by two analysts (the authors of this chapter – one whose L1 is Norwegian and one whose L1 is English) to weed out any constructions that were not PVs.

Running written learner texts through a PoS tagger such as CLAWS has possible drawbacks, because the tagger was developed on the basis of L1 English. While its success rate is high with regard to a standard variety of English, it might perform less optimally on L2 learner language, which is necessarily characterized by a higher degree of non-standard features. Because an earlier study into iden-

[3] Available from: http://ucrel.lancs.ac.uk/claws/

[4] Our database was created using the program *Filemaker Pro 12 Advanced*.

tification of lexical units in 20,000 words of the NICLE corpus, however, found CLAWS to be rather resilient in PV tagging, even for unconventional PVs (see Nacey 2013: 87), we decided that the PoS tagging of the entire NICLE corpus was not unwarranted.

Automatic tagging of the spoken transcripts of LINDSEI-NO is another matter entirely. Spoken language is characterized by a high number of disfluencies, including false starts, repetition, uncompleted sentences, overlap between speakers, as well as filled and unfilled pauses. All of these have been transcribed as faithfully as possible, making the transcriptions potentially problematic to tag automatically. As a result, we identified PVs in our spoken material by searching the corpus for all instances of prepositions and particles, and then sorted through all hits to manually separate particles in PVs from preposition use (or other use, such as the infinitive use of *to*). Only seven of these particles have no corresponding prepositional use: *aback, ahead, apart, astray, forth, forward*, and *together*. All sentences containing PVs were then added to our database for subsequent analysis of metaphoricity and conventionality.

The manual stage of PV identification was supplemented at times by dictionary consultation, particularly useful when distinguishing between PVs and free combinations. Here we relied on the online versions of two dictionaries intended for advanced learners of English: the *Macmillan English Dictionary for Advanced Learners* (MED) and the *Longman Dictionary of Contemporary English* (LM).[5] Although lexicographer standards for including a construction in learners' dictionaries as a PV are less stringent than those of Quirk et al. (1985), the inclusion of a phrase as a single lexical unit with at least one independent sense entry gives an indication that the phrase in question is often perceived as one semantic unit rather than two – including by language learners such as those who are targeted by our investigation.

3.2.1 Categorization for metaphoricity

All PVs were coded for metaphoricity following the Metaphor Identification Procedure Vrije Universiteit (MIPVU), a procedure developed for identification of metaphor in a reliable, replicable and theoretically valid way (see Steen et al. 2010). Following MIPVU, the contextual meaning of the PV is compared with its most basic sense found in the dictionary. If the two senses differ sufficiently, and

[5] Available from: http://www.macmillandictionary.com/ and: http://www.ldoceonline.com/

may be related by means of comparison (rather than generalization, specification, contiguity, etc.), the PV is marked as metaphorical. Otherwise, the PV is marked as non-metaphorical. MIPVU therefore requires a binary nominal yes/no decision as a means of ordering the reality of language into contrasting categories (Steen 2007: 93).

The basic sense is defined as the most concrete, specific and human-oriented sense, which is typically (though not always) the oldest sense. Note that the basic sense is not necessarily the most frequent sense, which is often listed as the first entry in frequency-based dictionaries produced for learners of English; indeed, a metaphorical sense is often the most frequent and/or salient sense. For a metaphorical mapping to exist, there must be a distinction between the basic sense and the contextual sense of the PV in question. Sufficient distinction is usually decided on the basis of lexicographical practice, whereby senses of a single lexical unit are listed as separate, numbered entries following the headword. This information provides evidence regarding sense distinction, enabling the analyst to judge whether the contrast between the various numbered entries for a single word is indeed sufficient to be compared for metaphor, i.e. to evaluate whether cross-domain similarity may be present. Such similarity may take one of several forms: "pre-existing as well as created similarity [...and...] literal or external similarity [or resemblance] as well as relational or proportional similarity (or analogy)" (Steen 2007: 63). Following the original MIPVU, we consulted MED and LM as our dictionaries, which are both written on the basis of corpus evidence, and provide illustrative sentences of the headword in context.

The PV *catch up with* in (3) provides an example from our data of how metaphors are identified following MIPVU.

(3) Our past eventually **catches up with** us, and then we will all have to face our wrongdoings. (AC5001)

While the contextual sense of this verb is "to begin to have an effect on someone" (2nd definition in MED), as in *The lack of sleep caught up with her, and she began to doze off*, its basic sense is the more concrete, specific and human-oriented sense: "to find and arrest someone who has committed a crime" (1st definition in MED), as in *The police will catch up with you sooner or later*. The relation between these two senses may be understood in terms of a comparison involving a concrete and abstract agent (e.g. *police* vs. *our past* / *the lack of sleep*). This occurrence is then coded as metaphor. By contrast, the PV *end up* in (4) and (5) is non-metaphorical:

(4) so: . that's why I **ended up** . here in Hamar (NO046)

(5) If your life is only filled with work, chaos, and pressure, you will **end up** sour and sad. (AC10001)

MED has only a single sense entry for this verb, "to be in a particular place or state after doing something or because of doing it", which is broad enough to correspond to the contextual meanings of both occurrences, despite the sense of (4), referring to a place, being more concrete than that of (5), referring to a state. Moreover, the fact that there is only a single sense entry necessarily entails that this particular PV will never be metaphorical in use, providing it is used in a conventionally codified sense. The contextual meaning will always correspond to the basic meaning, simply because there is no other codified sense with which there could be a contrast.

MIPVU considers PVs as single lexical units despite consisting of two or three elements, because they "function as linguistic units designating one action, process, state or relation in the referential dimension of the discourse" (Steen et al. 2010: 28). Put another way, evidence indicates that speakers "mentally lump [...] verb and particle together as a single word" (Lindstromberg 1998: 252). What this means is that any metaphorical reasoning underlying the choice of verb or particle of codified phrasal verbs was disregarded in the metaphorical categorization, as the PV represents only a single referent in the real world. This identification procedure, however, was developed on the basis of L1 English, from four different text types in the British National Corpus. Whether the operational decision to treat PVs as single lexical units is appropriate when it comes to L2 English is an issue. MacArthur (2014), for example, speculates on the degree to which such conflation of PV elements into a lexical whole may be regarded as a psychological reality for learners of English – and whether MIPVU should be modified as a consequence (see also MacArthur & Littlemore 2011: 210–211). This point is further discussed in 4.4.

3.2.2 Categorization for divergence and conventionality

As with the PV identification (Section 3.2) and the categorization for metaphoricity (Section 3.2.1), the analysis of learners' PV use according to divergence and conventionality relied on the learner dictionaries MED and LM, combined with manual interpretation.

Divergence, i.e. deviation from a standard, comprises different sub-categories in the data. In addition to fairly obvious learner errors such as spelling and subject-verb concord mistakes, divergence was registered in cases where we analyzed the choice of either the verb or the particle in the PV construction as inconsistent with the corresponding dictionary entry (including instances where a regular PV should be turned into a phrasal-prepositional verb or the other way around). Illustrative examples include *lock in* instead of *let in* in (6), *put up* instead of *put on [a play]* in (7) and *go through with* instead of *go through* in (8). Since the data was partly extracted by retrieving PV particles (see Section 3.3), instances where the particle may be missing have not been registered.

(6) when I got there I met . (eh) one of the students . he was kinda on guard that night he **locks** people **in** if they've locked themself out (NO031)

(7) so we're actually **putting up** Alice in Wonderland (NO016)

(8) We are given points on how to plan and **go through with** a lesson, and most subjects really does this well. (HO-0003.1)

Sometimes the entire PV selected by the learner represents a divergent choice contrasting with its dictionary entry. In such cases, a potential lexical target item which would normally be used by L1 language users may exist, as in (9) where *go off* seems to stand for the verb *close*.

(9) you know when the night club **goes off** then people start going to work (NO034)

One final consideration, especially crucial when working with learner language, is the possibility for neologisms – that is, PV-like constructions that are not codified in standard dictionaries of English. Gilquin (2015: 58) reasons that such “non-existing, deviant phrasal verbs [could be] *bona fide* constructions in the learner’s language system” and are consequently worthy of inclusion in an investigation of PV usage in learner corpus data. We agree with her reasoning, and have therefore chosen to retain all such constructions in the data. An example is *choose for* in (10), an expression that adheres to the identifying PV characteristic, but is not a codified lexical unit in any of the dictionaries consulted.

(10) I haven't been . (eh) sure about what to: **choose for** . or what to study but I i= (eh) last year I tried . law school (NO039)

Such PVs fall into a category of their own with respect to divergence, an ‘Other’ category. As they are not codified in dictionaries, there is no sense entry by which to contrast their contextual sense. Although some might be more contextually appropriate than others, we have chosen to not grade them along any sort of cline of appropriateness at this stage. Indeed, it is sometimes impossible to tell whether a particular PV resulted from error or intention. As an example, consider *wonder off* in (11).

(11) Letting the mind **wonder off** on its own can work as therapy. (AG1011)

This PV could be either the result of a spelling mistake (= *wander off*) or a neologism created by the learner. To avoid the omission of potential instances of creative learner usage, such PVs were considered as novel in the research data.

4 Corpus-based quantitative results

Here the results related to the research questions are presented, i.e. the occurrence of PVs in the corpora and the Norwegian L2 learners' production of divergent PVs, including the identification and analysis of contrasting patterns of PV usage across spoken and written modes, and whether divergent PV usage can be linked to metaphoricity and L1 transfer.

4.1 Phrasal verbs in LINDSEI-NO and ICLE

A total of 1,489 PVs were identified in the combined corpora: 1,051 PVs in NICLE and 438 in LINDSEI-NO. That there are 2.3 times more PVs in the written material can be explained to a large degree by the different sizes of the two corpora. NICLE consists of 317 texts with a word count of 211,725 whereas LINDSEI-NO consists of 50 texts containing 83,675 words in all: NICLE is thus 2.5 times larger than LINDSEI-NO. All told, the PVs include a combination of one of 200 verb types with one of 25 particle types, confirming that identification of PVs through their particles is a more effective method than searching for lexical verbs. Table 1 lists the particles found in our material, along with a rough approximation of the observed tokens:

Tab. 1: Observed occurrences of PV particles in NICLE and LINDSEI-NO.

Frequency	Particle
> 100	***up***, ***out***, *back*
51–100	*away*, ***down***, ***on***, ***off***, ***in***
25–50	*around*, *over*, *together*
10–24	*through*, *along*
< 10	*ahead*, *forward*, *apart*, *about*, *forth*, *aside*, *behind*, *by*, *into*, *astray*, *for*, *round*

Here it can be seen that only a handful of particles are very productive in forming PVs, with 11 of the 25 particles accounting for 25 or more occurrences; those particles presented in boldface script comprise Biber et al.'s (1999: 413) table of the six most productive English PV particles. Table 2 presents the PV lexical verbs in our data, also roughly divided by frequency. The largest frequency grouping, of verbs appearing fewer than 10 times, contains 161 different verbs – so many that we have chosen not to list them all.

Tab. 2: Observed occurrences of PV verbal elements in NICLE and LINDSEI-NO.

Frequency	Lexical verb
> 100	***go***
51–100	***get***, ***come***, ***take***, *end*, ***turn***
25–50	*grow*, ***put***, *find*, *keep*, ***bring***, *sit*, *lock*, *look*, *make*
10–24	*give*, *point*, *figure*, *call*, *pick*, *move*, *pay*, *carry*, *run*, *send*, ***set***, *start*, *try*, *build*, *fill*, *show*, *work*, *dream*, *sum*, *walk*, *hang*, *cut*, *hold*, *wake*
< 10	161 verbs

The most productive verbs are relatively few, and to some extent match Biber et al.'s (1999: 413) list of the eight common PV verbs (indicated in boldface script in Table 2). Such frequency testifies to the extreme polysemous nature of these lexical verbs, which appear in combination with a variety of particles to create multiword verbs.

Finally, the 1,489 PV tokens are realized by 457 PV types, 373 in ICLE and 180 in LINDSEI-NO. Table 3 presents these PVs, along with an approximation of their observed frequency.

Tab. 3: Observed occurrences of PVs in NICLE and LINDSEI-NO.

Frequency	Phrasal verb
> 50	*end up*
40–49	*go on, grow up*
30–39	*find out*
20–29	*go out, turn out, lock up, make up, come up with, point out, sit down, figure out*
10–19	*go back, take away, go back to, pick up, come back, take over, get away with, try out, build up, get back, get up, keep up, come out, sum up, work out, show off, carry out, give up, keep on, take up, turn around, wake up*
< 10	423 phrasal verbs

Here we see that most PVs are best characterized by their infrequency, not unexpected given the Zipfian distribution of elements in slots that we find everywhere in corpora; the vast majority appear fewer than 10 times in the two combined corpora (and being so numerous, they are not specified here). That the written texts contain more PV types than the spoken texts is partly a consequence of the text topics. The LINDSEI-NO interviews are rather homogenous, the main topics being the students themselves: their hobbies, travels, education, etc. By contrast, the NICLE essays deal with one of 20 different topics – ranging from breakfast to whether money is the root of all evil – and hence call for PVs expressing a wider variety of meanings. For example, the PV *lock up* appears only in NICLE (24 times), triggered by two topics related to crime and prison system – a topic never raised during the LINDSEI-NO interviews.

The observed occurrences of PVs in the two corpora are equivalent to a relative frequency of 50.4 PVs per 10,000 words. This type of number is not very informative in and of itself, and comparison is therefore helpful to gain some sense of its significance. The one study that most closely parallels our own is Gilquin (2015), where all PVs in both native-speaker Louvain corpora, LOCNESS and LOCNEC (155,167 and 118,398 words respectively), were identified, as well as all PVs in the French subcorpora of LINDSEI (LINDSEI-FR, 91,440 words) and ICLE (ICLE-FR, 190,544 words). In her study, PVs were also identified through their particles, although Gilquin restricted herself to a concordance search for 24 particles from Huddleston and Pullum's (2002: 281) list of the 25 most central particles in the 'verb-particle-object-construction'. Our search was thus far more comprehensive, yet resulted in the identification of only six PVs that would not have been uncovered by Gilquin's search: two PVs each with the particles *behind* and *into*, one with *astray* and one with *for*. Consequently, the relative PV frequencies uncovered in the two investigations are highly comparable: these are presented graphically in Figure 1 where Gilquin's reported frequencies (2015: 62) are reproduced and ours are added.

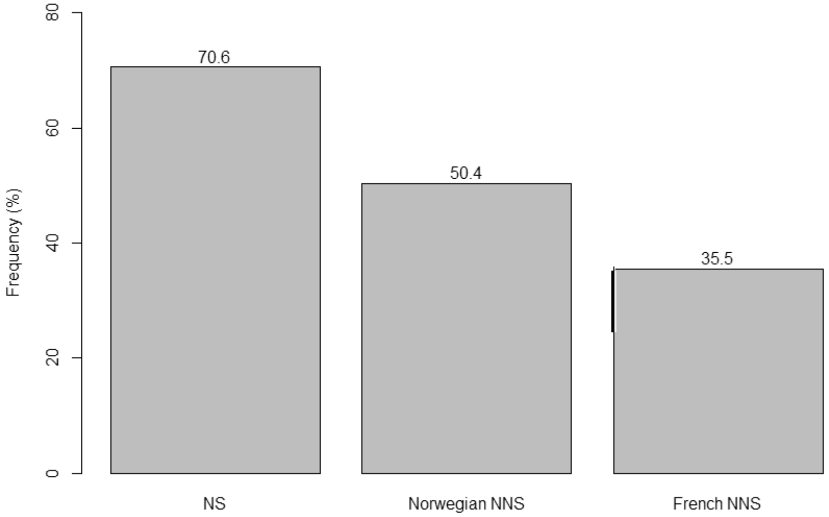


Fig. 1: Relative PV frequency per 10,000 words in spoken and written native speaker (NS) and non-native speaker (NNS) texts.[6]

On the basis of her data, Gilquin concludes that the French learners underuse PVs when compared to L1 speakers of English, something observed in other learner populations as well, such as in Chinese and Malaysian L2 speakers of English (see Liao & Fukuya 2004; Kamarudin 2013). A decisive factor for PV frequency in L2 English would seem to be whether or not the learners' L1 has a PV construction; French does not, leading to a possible explanation for the relative underuse uncovered by Gilquin. Adding the Norwegian data to that provided by Gilquin lends support to the importance of the L1. The relative PV frequency in Norwegian L2 English is neatly sandwiched between the NS group and the French NNS learners; this indicates that while Norwegians are a learner population and thus may not

[6] All graphical visualizations and statistical tests of significance were carried out using R (see R Core Team 2015 in the references).

use PVs as consistently or often as L1 English speakers, their PV usage is nevertheless significantly greater than that of the French, due to Norwegian being a language with a phrasal verb construction. In addition, the fairly high PV frequency might also be an indication of a generally higher English L2 proficiency in the Norwegian population, compared to many other populations (cf. EF Education First 2016).

Gilquin (2015: 63) continues by noting that the overall figures conflating spoken and written usage conceal important variation. Specifically, while the native speakers use far more PVs in speech than in writing, the French speakers do the opposite. Her figures are reproduced in Figure 2, with the comparable figures from the present investigation added.

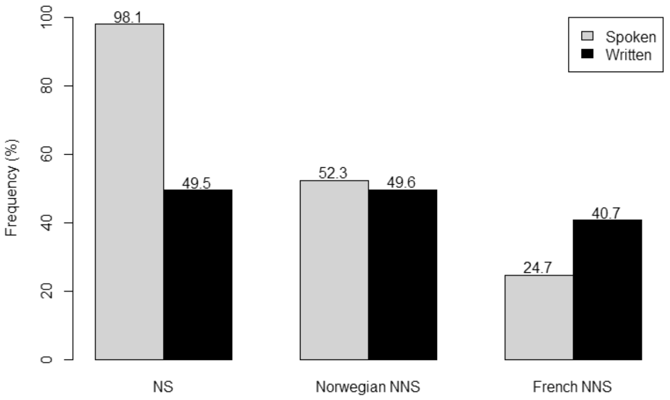


Fig. 2: Relative PV frequency per 10,000 words in NS and NNS (spoken versus written) texts.

Unlike both the NS and the French NNS groups, the Norwegians show no significant differences in PV frequency across the spoken and written modes. Further, Figure 2 indicates that PV frequency in written NS English and written Norwegian L2 English closely correspond. The relative PV underuse by Norwegian learners when compared to L1 English speakers may therefore be due to an underuse in their *spoken* English.

A danger with calculations such as those represented in Figure 2, however, lies in their being based on aggregate data that essentially treats each corpus as if the texts were uniform. Individual variation is neglected. To allay such doubts regarding our conclusions in this area, we also explored the PV frequency per text and mode, visualized in the boxplot in Figure 3.

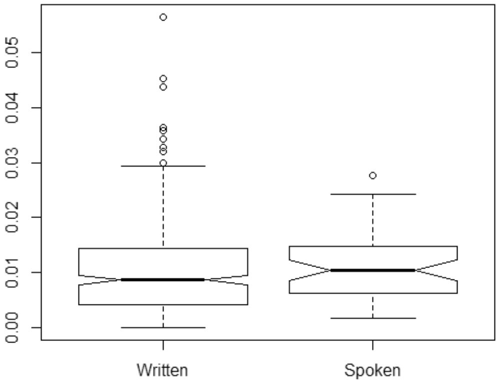


Fig. 3: Observed PV occurrences (%) per word total and text, spoken versus written.

While every LINDSEI-NO text has at least one PV and a Shapiro-Wilk test indicates that the spoken texts are normally distributed (W = 0.96024, *p* 0.0910), we find that the written material is right-skewed (W = 0.89923, *p* < 0.0001). There are more exceptionally small values (39 of the 317 texts have no PVs at all), and there are also a few exceptionally large values. A Mann-Whitney-Wilcoxon test, however, fails to show any significant differences in population distribution across the spoken and written modes (W = 7028.5, *p* 0.1985). This indicates that the Norwegian spoken and written texts may behave similarly with respect to PV frequency, unlike the NS and French NNS texts in Gilquin's (2015) investigation.

4.2 How often do learners produce divergent phrasal verbs?

Roughly one in ten of all PVs in both NICLE and LINDSEI-NO are divergent, the divergence rates being 9.9% and 10.7% respectively. Text mode thus plays no role in terms of the frequency of PV divergence.

The majority of the divergent PVs are related to structural problems, such as (12), where a single verb, *listed*, would be more contextually appropriate (see also examples (6) – (8) in Section 3.3.2).

(12) It is therefore difficult to say that all the positive characteristics of the Americans that Turner has **listed up** in his article are describing the American people. (UO-0005.1)

In several instances, the PV is semantically related to an appropriate target verb, but does not comply with the collocational range of the target language PV, as in (13) where the proposed verb could be *trigger* or *stimulate*. In some cases the learners also use more general core lexemes than what is expected (cf. Hasselgren 1994), such as in (14), where a more appropriate verb could be *install*.

(13) [...] the impressions and impulses that meet you in your everyday-life can indeed work as keys to **start off** your imagination. (BE-0010.1)

(14) I think my . parents were the ones to . in our younger days .. just **get** water **in** . but no I guess not (NO013)

4.3 Are there contrasting patterns of phrasal verb usage across the spoken and written modes?

Even though there are no significant differences between the corpora with respect to divergence frequency, differences do become apparent when it comes to the distribution of divergent PVs across texts. Figure 4 shows the percentage of divergent PVs per text and per mode.

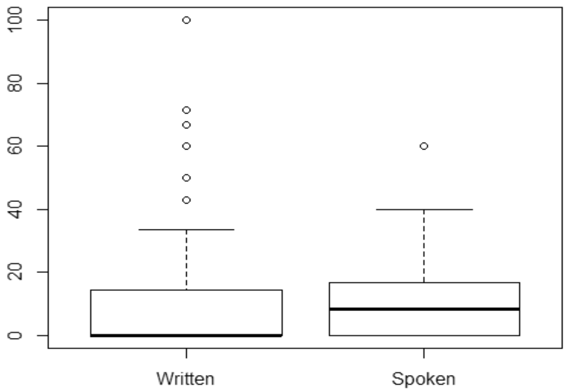


Fig. 4: Percentage of divergent PVs per text and mode.

Here we see that the median for the number of divergent PVs in the written texts is zero. This is because 39 of the 317 total texts have no PVs at all (and hence no divergent ones), while 198 of the remaining 278 texts contain only PVs whose contextual sense matches a conventionally codified sense and/or novel PVs which have been classified as 'Other' in terms of divergence (see Section 3.2.2). Only 25.2% of the NICLE texts contain any divergent PVs. The median for spoken texts is higher, although the LINDSEI-NO boxplot lacks a left whisker because there are some texts with no divergent PVs. Although each of the 50 LINDSEI-NO interviews contains at least one PV, only 28 of them (56%) contain one or more divergent PVs. The differences indicated here are significant (W = 5411.5, $p \approx 0.003108$).

Finally, Figure 5 shows the same data as in the previous diagram, but only for those 81 NICLE texts and 28 LINDSEI-NO texts that contain at least one divergent PV, a concern being that the abundance of texts with either no PVs or no divergent PVs perhaps conceals important information.

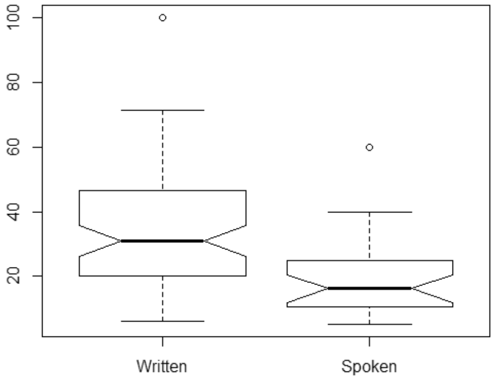


Fig. 5: Percentage of divergent PVs per text and mode, ONLY for texts with divergent PVs.

When texts containing no divergent PVs are disregarded, we see that the median for divergent PVs is higher in the written texts than in the spoken texts. Adding the information gleaned from Figure 4 and Figure 5 together, we conclude that there are contrasting patterns of PV usage across the written and spoken modes. In general, more learners overall produce divergent PVs when they speak than when they write (as indicated in Figure 4), ostensibly due to greater pressure on processing time; editing is more difficult in conversations. Indeed, our LINDSEI-NO informants rarely corrected divergent PVs themselves (nor were they corrected by the interviewer), but instead continued with their intended message. Problems with PVs depend more upon individual differences when it comes to written language, whereas they seem to be a more general characteristic of speaking. There is greater opportunity for text editing in the written texts, which may explain the far greater proportion of NICLE texts with no divergent PVs. Figure 5 adds nuance to this picture, however, by showing that in the (fewer) cases when learners *do* write divergent PVs, they produce a higher percentage of them than in spoken texts where learners have uttered divergent PVs. In other words, divergent PVs in written texts are more likely to indicate a real language gap, while in spoken texts they may just indicate a temporary slip.

4.4 Is there a link between divergent phrasal verb usage and metaphorical usage?

Based on the MIPVU identification procedure, slightly more than one-third of the PVs in our collected data, 564 PVs (37.8%), are metaphorical, while 925 PVs (62.1%) are not metaphorical. These figures may be compared with those from previous studies using the same procedure for metaphor identification. Steen et al. (2010: 780), investigating overall metaphorical frequencies in four different genres from the Baby-BNC, report frequencies ranging from 6.8% for conversation to 17.5% for academic texts. Nacey (2013: 136), who used MIPVU to identify all metaphors in roughly 20,000 words of text in both written L1 English and Norwegian L2 English (from the NICLE corpus, just as in the present study), reports a metaphor frequency of 16.7% and 18% respectively. From this, one may conclude that although the majority of PVs in our data are non-metaphorical, the proportion of metaphorical PVs is nevertheless considerably higher than the metaphor/non-metaphor proportion of language in general, both in L1 and L2 English.

When it comes to the proportion of divergent metaphorical and non-metaphorical PVs, we find that 51 (5.5%) of the 925 non-metaphorical PVs display semantic or syntactic inappropriateness, while 75 (13.2%) of the 564 metaphorical PVs are divergent. Although this difference is intriguing, we would hesitate to claim any negative correlation between metaphoricity and conventionality without further detailed investigation. More specifically, these numbers reflect the aggregate data of 25 particle types, 200 verb types, and 457 different PVs, either written by one of 317 authors or spoken by one of 50 interviewees. Uncovering the link (if any) between conventional appropriateness and metaphoricity would require mapping the correlation between each particle, verb, and PV with respect to conventionality and metaphoricity per text and medium to discover which factors are the most statistically significant. This task would arguably require a much larger corpus to generate meaningful numbers, and therefore remains an open question for future research. In this respect, our data is underpowered.

A closer look at the 76 novel PVs in our data – that is, those that have been categorized as novel since there are no dictionary entries for them – offers an alternative approach towards investigating possible links between metaphoricity and divergence. Looking again at *end up* in (4) and (5) in Section 3.2.1 (reproduced below for the sake of convenience), we may note that the particle *up* is metaphorical in some sense.

(4) so: . that's why I **ended up** . here in Hamar /(NO046)
(5) If your life is only filled with work, chaos, and pressure, you will **end up** sour and sad. (AC10001)

The reason why we say *end up* rather than *end down*, *end in*, etc. is that *up* reflects a sense of completion and/or thoroughness (a metaphorical extension of the physical [location] sense of *up*), matching the sense conveyed by perfective verbs (see Lindstromberg 1998: 24). Due to the figurative extension of the meaning of the particle, the PV is metaphorical in origin, even though it is not metaphorical in use when analyzed for metaphoricity on the basis of the complete PV.

Novel PVs, by contrast, cannot be analyzed for metaphor as single lexical units, as they are per definition not codified in dictionaries. Because there is no entry for the PV as a whole, we must rely on the entries for the individual elements to judge metaphoricity. This forces us to deviate from the standard MIPVU procedure for identifying metaphorical PVs, where each PV is treated as a single lexical unit. This procedural necessity, in turn, provides an opportunity to evaluate the effects of deviating from the standard MIPVU guidelines by analyzing PVs in learner language differently from how they are treated in native-speaker language. The procedural deviation here is to categorize novel PVs as metaphorical in use if *either* the verb *or* particle involved (or both) is judged metaphorical after dictionary consultation. Following this procedure, we find that 60 (78.9%) of the 76 novel PVs in our data are metaphorical, while 16 (21.1%) are not. This means that at least one of the constituent elements of most novel PVs is metaphorical, either the lexical verb or the particle.

Looking more closely at the individual novel PVs in our data shows that while the verb is sometimes metaphorical, the particle nearly always is metaphorical. As an example, consider the particle *away*. It is the most frequent particle among the novel PVs in our data, found in the following PVs: *dream away* (8 occurrences across 7 different texts), *babble away, choose away, pluck away, promise away, learn away, scratch away, drift away* (2 occurrences in separate texts), *fly away, shuffle away*, and *glide away*. Of these PVs, all but two are metaphorical; *scratch away* is non-metaphorical, used in reference to a cat scratching all his fur away

(LINDSEI, NO006), as is *pluck away* referring to doctors removing bad genes (NICLE, UO1009). Among the others, all instances of *away* are metaphorical, whereas the only metaphorically-used verbs are those denoting physical movement (that is, *drift, fly, shuffle* and *glide* used in a context calling for figurative movement, as when *thoughts drift away* [UO10007]).

Moreover, we find that the Norwegian learners tend to use particles in ways that are wholly consistent with their conventional figurative extensions, as illustrated by the PVs with *away* in (15) and (16):

(15) most people, no matter how busy they are, spend some time to "**fly away**" from all worries in the real world. (BE-0002.1)

(16) I think they have to just lie down on the couch and **dream away** every once in a while. (OS-0041.1)

The basic sense of *away* is the opposite of *toward*. It is neutral with respect to both the starting and ending points. About its use in PVs in particular, Lindstromberg (1998: 257) adds the following:

> *Away* is fairly common in phrasal verbs. [...], it typically contributes the meaning of 'without end' (e.g. *while away the hours*) or, by extension 'with abandon', as in *He gaily whistled away*. This is entirely consistent with *away*'s latent lack of end-point focus.

Following this reasoning, *fly away* in (15) exemplifies the primary metaphorical extension of the particle, referring to an action in terms of metaphorical movement along an unspecified path away from a point of reference (here, *worries in the real world*). This particular example, found in the written material, also has square quotes encasing the PV, presumably marking the writer's awareness that said *flying away* is not literal (see Nacey 2012). *Away* in (16) denotes 'without end', having no end point. *Dream away* is the most frequent novel PV in our material, occurring eight times across seven texts, and is found among the texts written in response to a prompt about dreaming and imagination. An additional contributory factor in its production is L1 transfer, as Norwegian has a closely corresponding reflexive PV *drømme seg bort* (lit.: 'dream oneself away', colloq.: 'daydream'; see also Section 4.5).

Other novel PVs contain what is in essence a superfluous particle where the meaning of the particle is appropriate, but encoded in the verb at hand. We see this in the aforementioned PV *spend away* (= 'spend'), as well as in others such as *cover over* (= 'cover'), *fill over* (= 'fill'), *sync together* (= 'sync'), and *end off* (= 'end'). Other novel PVs in our data follow patterns established by codified PVs. An example is the PV *rush off* in NICLE, not found in dictionaries even though the comparable PV *dash off* is listed in dictionaries. This illustrates the limitations of our method relying on dictionary codification, as no dictionary is able to contain all lexical items even though they might be evident in use.

In sum, the vast majority of novel PVs in the data adhere to patterns evident in codified metaphorical extensions, albeit with varying degrees of idiomaticity – from the awkward and syntactically incorrect *comment back to* (for 'reply') in (17) to the very idiomatic *drive apart* in (18). In only a single instance is the choice of particle completely inappropriate in that it conveys the opposite sense of the verb: *bring apart* in (19), where the writer varies the lexis through altering the particle rather than the verb.

(17) Gabriel blushes and feels he had made a mistake. He does not **comment back to**, he tries to pretend like nothing by kicking of his goloshes. (UO1002)

(18) One can say that the absense of dreams and imagination **drive** people **apart**. (UO1085)

(19) Instead of **bringing** us **apart**, this technology has brought us together. (HO10029)

One reason offered for the purported difficulties L2 learners of English may have with PV is their polysemy, and in particular their (sometimes several) figurative extensions. We have not investigated the link between conventional appropriateness and metaphoricity beyond presenting aggregate figures and thus cannot confirm or reject this supposition on the basis of our material. What we have found, however, is that the Norwegian learners represented in our data are adept in creating novel metaphors to convey meaning in much the same way as native speakers, through established metaphorical extensions of particles. In other words, metaphor is an aid rather than a barrier to communication, both in speech and writing.

4.5 Is there a link between divergent phrasal verb usage and L1 transfer?

According to Waibel (2007: 122), it is "obvious that the influence of the native language has considerable impact on learners' performance" related to PV use. Specific positive L1 transfer is hard to detect; however, comparisons of the PV frequency in learner corpora with different L1s clearly reflect L1 influence. An example from Waibel (2007: 85) is learners with a Germanic L1 who have the same frequency of PV production in English L2 as native speakers of English, in contrast to learners with other L1s (Romance languages, Japanese, etc.), whose frequency of PV production is much lower. Since Norwegian is a Germanic language, the PV pattern is already inherently in place, as indicated in Section 4.1, Figure 1 and 2. In addition, many basic verbs and particles with a high frequency are both formally and semantically equivalent in Norwegian and English, e.g. *drink up – drikke opp* and *fill out – fylle ut*.

For the same reason, negative L1 transfer is most easily detectable in corpus data from learners whose L1 has similar verb-particle constructions as in English. A number of the divergent PVs in the material have Norwegian structural equivalents which may have affected the learners' choice, such as (20), where the student is talking about teaching practice, which in Norwegian may be expressed as *gå ut i praksis* (lit.: 'go out in praxis').

(20) but (em) .. I'm looking forward to: . **going out on** . **in** the next . practice period (NO039)

One PV element where L1 transfer appears to be relevant is the particle *away*, which is also the most frequent particle in novel PVs in the corpus data (see Section 4.4). A majority of these occurrences correspond to Norwegian PVs with semantically equivalent verbs and particles (English *away* = Norwegian *bort*), e.g. *dream away – drømme bort* ('daydream; get lost in a dream'), *choose away – velge bort* ('choose instead of'), *promise away – love bort* ('commit/oblige'), *learn away – lære bort* ('teach'), etc. Such examples indicate that direct PV translation may sometimes seem like the most natural choice for learners. Since the probable English lexical target items in most of these instances are not PVs, and/or do not contain the particle *away*, they are among the most likely cases that show potential transfer. However, for most of the divergent PVs no clear lexically and/or semantically equivalent Norwegian PVs can be detected, which means that even potential influence of the learners' PV use is difficult to determine. This emphasizes that reliable research on L1 influence requires large amounts of data, and triangulation including, for instance, experimental research and comparative research of consistent learners with different L1s, which is beyond the scope of the present study.

5 Conclusions

In this chapter, we have presented corpus-based data that contributes to the knowledge of PV use by learners of English, by investigating all PVs in two Norwegian learner corpora, NICLE (written) and LINDSEI-NO (spoken). The reason for choosing PVs as the focus of investigation is that learners' mastering of English PVs is generally described as both difficult and problematic, as well as highly important for their English language proficiency.

Our findings challenge the general perception of PVs as a highly problematic linguistic structure for language learners. The Norwegian learners' PV use both in spoken and written mode matches the frequency rate of the comparative native speaker written corpus, around 50 per 10,000 words (although the frequency in NS speech is higher). Divergent usage occurs only in about 10% of the cases (which are not all necessarily incorrect; see Section 3.2.2), and here, too, the same rate occurs both in speech and writing. In many respects, our findings are also in alignment with some previous studies based on the same corpus collection from CECL (Waibel 2007; Mondor 2008; Gilquin 2015), which show that the learners' L1 affects the frequency of PV use in learner corpora. Since Germanic languages, including Norwegian, have PVs, it will be natural for learners to make use of this structure in their foreign language learning (cf. Waibel 2007).

In order to confirm whether the spoken and written modes actually are as similar as the corpus material indicates, the findings related to divergent PVs in the present study were also examined by looking beyond the aggregate data to consider the learners' individual use. By adding this statistical perspective, significant differences appeared which show that while several written texts contain no PVs, those learners who produce divergent PVs in writing often have a relatively high number. On the other hand, all the spoken texts do contain PVs, but the divergent PVs in the spoken material are more evenly spread out. Whether this is a more general tendency in similar corpora produced by learners with other L1s is a question for future research, since at the time there are no parallel studies as far as we know.

Concerning metaphoricity, the results of the present investigation reveal that the proportion of metaphorical PVs in the corpus data is considerably higher than the proportion of metaphor use in language in general. Metaphorical use of PVs is also believed to present a problem for learners, a claim that ostensibly calls for studies of correlation between figurative use and divergent PVs. We question the empirical basis for the allegation that metaphorical PVs pose difficulties for L2 language learners, as looking into this research question requires a great deal more data than currently exists. It also requires a reliable means of identifying metaphoricity, something that metaphor researchers have only started taking seriously in recent years (see e.g. Nacey 2013; Steen et al. 2010). In the absence of any clearly explained metaphor identification procedure, it becomes impossible to compare results across studies that make claims about PVs and figurative use.

That said, while this study employed the Metaphor Identification Procedure Vrije Universiteit, our exploration of the novel PVs in the learner corpus data demonstrates that valuable insight may also be gleaned from analyzing the metaphorical status of the individual elements of PVs, rather than following standard MIPVU practice that considers PVs as single lexical units. In the case of these advanced Norwegian learners of English, for instance, we argue that metaphor might provide more help than hindrance, as they appear to use conventional metaphorical extensions of particle meanings to communicate meaning through novel PVs in much the same way as native speakers.

Potential negative L1 transfer might play a significant role on the use of divergent PVs, but, as mentioned, only a small number of the divergent PVs in the data indicate a direct semantic relationship with corresponding Norwegian PVs. The use of one specific particle, *away*, appears as a fairly plausible result of L1 transfer, and since it is not mentioned in any language guidebooks that we can find, it might imply some consequences for the teaching of English, and should also inspire further investigation of individual lexical items.

Finally, the results from this study indicate a need for more commensurate findings from other studies examining the use of PVs by other groups of language learners, e.g. a comparison of learners with different L1s (or only different Germanic L1s). Since the present study is limited to advanced learners, investigations into the use of PVs among younger and/or less proficient learners would also be valuable, to add to the present findings. In short, there is still work to be carried out.

References

Askedal, J. O. (1994). Norwegian. In E. König & J. v. d. Auwera (Eds.), *The Germanic Languages* (pp. 219–270). London, UK: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Cowie, A. P. (1993). Getting to grips with phrasal verbs. *English Today*, *9*(4), 38–41.

EF Education First. (2016). *EF English Proficiency Index* (6th edition). Retrieved from http://www.ef.no/epi/ (last accessed March 2017).

Gilquin, G. (2015). The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory*, *11*(1), 51–88.

Gilquin, G., De Cock, S., & Granger, S. (2010). *Louvain International Database of Spoken English Interlanguage (LINDSEI)*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–260.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance. Grammatical or semantic causes? *Studies in Second Language Acquisition*, *11*(3), 241–255.

Jenkins, J. (2009). *World Englishes: A Resource Book for Students*. London, UK: Routledge.

Kamarudin, R. (2013). A Study on the Use of Phrasal Verbs by Malaysian Learners of English. Doctoral dissertation, University of Birmingham, Birmingham, UK.

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning? L1–L2 difference, L1–L2 similarity, or L2 complexity? *Studies in Second Language Acquisition*, *15*(1), 35–48.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, *54*(2), 193–226.

Lindstromberg, S. (1998). *English prepositions explained*. Amsterdam, Netherlands: John Benjamins.

MacArthur, F. (2014). Metaphor and the learner of English as an international language. Paper presented at the conference *Researching and Applying Metaphor (RaAM) 10: Metaphor and Communication in Science and Education*. June 20–23. University of Cagliari, Cagliari, Italy.

MacArthur, F., & Littlemore, J. (2011). On the repetition of words with the potential for metaphoric extension in conversations between native and non-native speakers of English. *Metaphor and the Social World*, *1*(2), 201–238.

Mondor, M. (2008). *Figuring* it *out*. A Corpus-Based Comparison of the Verb-Particle Construction in Argumentative Writing by Swedish Advanced Learners and Native Speakers of English. Doctoral dissertation, University of Gothenburg, Gothenburg, Sweden.

Nacey, S. (2012). Scare quotes in Norwegian L2 English and British English. In S. Hoffman, P. Rayson, & G. Leech (Eds.), *Corpus Linguistics: Looking Back, Moving Forward* (pp. 117–130). Amsterdam, Netherlands: Rodopi.

Nacey, S. (2013). *Metaphors in Learner English*. Amsterdam, Netherlands: John Benjamins.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London, UK: Longman.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (last accessed April 2017).

Steen, G. J. (2007). *Finding Metaphor in Grammar and Usage*. Amsterdam, Netherlands: John Benjamins.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam, Netherlands: John Benjamins.

Waibel, B. (2007). Phrasal Verbs in Learner English: A Corpus-based Study of German and Italian Students. Doctoral dissertation, Albert-Ludwigs-Universität, Freiburg, Germany.

Keiko Tsuchiya

Conversational gesture corpus analysis

A method to analyse the strategic use of learners' gestures in paired English conversations

Abstract: The purpose of this study is to establish the method of conversational gesture corpus analysis (CGCA), which integrates multimodal corpus linguistics (MCL) with conversation analysis (CA), to investigate Japanese learners' strategic use of gestures in English conversations, especially in 'repair' sequences. The study compares how Japanese advanced learners of English and basic-level learners use hand gestures in pair conversations. CGCA was applied to investigate: (1) the word count and time lengths of speaker turns, (2) the frequency and functions of hand gestures, and (3) the use of gestures in repair sequences, comparing the two levels. Some differences were observed, i.e. the advanced learners self-repaired with metaphoric gestures, while the basic learners other-repaired with iconic gestures. The method made it possible to gain an overview of a global pattern of the temporal relationship between speech and gestures from which specific cases were selected for micro-analysis.

1 Introduction

Spoken corpus linguistics has two subcategories: monomodal spoken corpus analysis, focusing on "spoken language description", and multimodal spoken corpus analysis, which examines "alignments between language and hand gestures" in reference to other modes, such as head nods and prosody (Adolphs & Carter 2013: 2). This chapter addresses the latter. Knight (2011: 204) defines a multimodal corpus as:

> a linguistic corpus which presents records of interaction in different data streams within the corpus interface, integrating and aligning video and audio records alongside the more traditional text-based transcripts.

The approach of multimodal corpus linguistics (MCL) has been applied in recent spoken corpus studies (Adolphs & Carter 2013; Adolphs, Knight & Carter 2011;

Keiko Tsuchiya, Yokohama City University, ktsuchiy@yokohama-cu.ac.jp

https://doi.org/10.1515/9783110489071-017

Baldry & Thibault 2006; Knight 2011; Knight et al. 2009; Tsuchiya 2013). Baldry and Thibault (2006) develop a multimodal concordancing system to analyse gazing in TV car advertisements. Knight et al. (2009) investigate depths of head nods in relation to the degree of agreement. Adolphs, Knight and Carter (2011) report their experiments on 'heterogeneous' and 'ubiquitous' multimodal corpora, which include data from SMS (short messaging service) and GPS (global positioning system) to capture everyday communication. Knight (2011) focuses on the collocations between gestures and utterances in conversation.

Similarly to Knight (2011), Tsuchiya (2013) investigates the temporality of the use of hand gestures and head nods in listenership in reference to turn-taking patterns in intercultural encounters. For the analysis, a time-aligned corpus was developed, which aligns utterances and two other modes, hand gestures and head nods, on a timeline (see Section 3 for the detail of the methodology). On the basis of Tsuchiya (2013), the current study aims to establish the methodology of a conversational gesture corpus analysis (CGCA), to which both the temporality and the types of hand gestures are annotated in alignment with spoken language for the multimodal analysis of the learner interactions. A small-scale multimodal corpus was developed to compare hand gestures of advanced-level Japanese learners of English in a pair conversation with basic-level counterparts. Through the observation of the current data, I decided to focus on hand gestures in 'repair' sequences. Repair is defined by Schegloff (2007: 101) as "efforts to deal with trouble-sources or repairables – marked off as distinct within the ongoing talk" (see Section 2.2 for the review of repair strategies) and a central concept in ethnomethodological conversation analysis (CA). The methods of CA enable researchers to "describe the underlying social organization – conceived as an institutionalized substratum of interactional rules, procedures, and conventions – through which orderly and intelligible social interaction is made possible" (C. Goodwin & Heritage 1990: 283).

The main purpose of this study is to establish the research methodology of CGCA, which integrates MCL with CA, to investigate how Japanese learners strategically use gestures in a dyad English conversation, especially in repair sequences. I assume that Japanese learners of English with different English proficiency levels might use gestures with different functions to establish and maintain the interactions. Three research questions are addressed here: (1) are there any differences in the word count and length of speaker turns between the two pairs?, (2) how many and what kind of hand gestures do the learners use in the conversations, and are there any differences between the two levels?, and (3) how do these hand gestures relate to their repair strategies? The size of the corpus for the case study is quite small and the number of words in each conversation is less than 500. However, it contains the descriptions of the use of hand gestures and repair strategies, which are aligned with spoken language on a timeline. This corpus design allows the researcher to conduct both quantitative and qualitative analyses of the interactions.

Multimodality has also been of interest to discourse analysts since the 1980s (C. Goodwin 1981; M. Goodwin & C. Goodwin 1986; Jewitt & Jones 2008; Kress 2011; Norris 2008; O'Halloran 2011; van Leeuwen & Jewitt 2001). Recent studies include analyses of gaze, posture and gesture produced by: a teacher in the classroom (Jewitt & Jones 2008); an owner-worker interaction at the office (Norris 2008); and politicians' talks in a television debate (O'Halloran 2011). Three different approaches in multimodal discourse analysis are summarised by Jewitt (2009: 8–13): social semiotic multimodality (Kress 2011; Kress & van Leeuwen 2001; van Leeuwen 2005), systemic-functional multimodal discourse analysis (O'Halloran 2008; O'Halloran 2011; O'Toole 1990), and multimodal interactional analysis (C. Goodwin 1981; M. Goodwin & C. Goodwin 1986). Social semiotic multimodality places emphasis on a sign-maker as a social actor, while systemic-functional multimodal discourse analysis derives from Halliday's (1978) systemic functional linguistics and focuses on "the meaning potential of semiotic resources" in a particular context (O'Halloran 2008: 444). Multimodal interactional analysis maintains Goffman's (1955) social interactional theory. This study fits in with the third approach since how learners represent themselves in the interactions is considered in the discussion (see Section 5.2).

2 Describing gestures in conversation

This section provides a literature review on two themes relevant to the present study: how hand gestures have been examined in terms of their temporality with words, and how hand gestures in repair sequences have been described in existing studies.

2.1 Gesture – its categorisation and temporal relationship with words

Previous studies on hand gestures, especially focusing on their temporal relationship with speech sounds will be reviewed in this sub-section. There are many ways to categorise gestures, but three frameworks in McNeill (1992) and more recent studies of learners' gestures by Alibali and her colleagues (Alibali & Kita

2010; Alibali & Nathan 2012; Goldin-Meadow & Alibali 2013) are particularly relevant to the present study.

Based on the definitions of manual movements in talk in McNeill (1992), Goldin-Meadow (1999: 422) describes four functions of hand gestures:

1. 'Iconic' gestures transparently capture aspects of the semantic content of speech.
2. 'Metaphoric' gestures are like iconics in that they are pictorial; however, the pictorial content is abstract rather than concrete.
3. 'Beat' gestures look as though they are beating musical time.
4. 'Deictic' or pointing gestures indicate entities in the conversational space, but they can also be used even when there is nothing to point at.

Using this categorisation, Adolphs and Carter (2013), for example, identified the use of gestures in a particular context, i.e. the frequent use of beat gestures in lectures.

McNeill (1992) distinguished two types of iconic gestures which reflect a narrator's different viewpoints: a 'character view point' (C-VPT) gesture, which "incorporates the speaker's body into the gesture space" (McNeill 1992: 119), and an 'observer view point' (O-VPT) gesture, which "excludes the speaker's body from the gesture space and his hands play the part of the character as a whole" (McNeill 1992: 119). Based on McNeill (1992), Beattie, Webster and Ross (2010) focused on the relationship between listeners' attention and speakers' use of gestures, analysing listeners' fixations of eye movements in relation to speakers' use of shorter/longer C-VPT gestures and O-VPT gestures, finding that fixation was observed more frequently when the speaker used a shorter C-VPT gesture, which, thus, is considered to be more effective and communicative than other forms of gestures. In the analysis of the current study, the two types of iconic gestures, C-VPT and O-VPT, and the use of gestures to construct interactions between a speaker and a listener are considered.

In terms of the process of producing a gesture, within a single movement of a hand gesture, McNeill (1992: 25) defined three phases based on Kendon (1980): 'onset' ("preparation for the gesture"), 'stroke' ("the main part of the gesture") and 'retraction' ("the return of the hand to quiescence"). Stroke is obligatory while the other two phases, onset and retraction, are optional, depending on the position of the interlocutor's hands at the commencement of the movement. The temporal relationship between speech and accompanying speech gestures have been one of the central issues in the discussions in psychology. Among the previous studies of pragmatic and cognitive relations between gesture and speech, Butterworth and Hadar (1989) report the relationship between word retrieval and an iconic gesture. Chui's (2005) analysis of the use of iconic gestures in Chinese conversations revealed that hand movements tend to start prior to the 'affiliated words' in non-fluent speech, but the percentage of the occurrences decreases in fluent speech. More recently, Loehr (2012) examined the synchrony of gesture and intonation, finding the alignment of gesture with tones and intermediate phrases, which, thus, indicates a close relationship between gesture and cognitive/pragmatic structures.

The roles of gestures in children's learning stages have also been investigated (Goldin-Meadow & Alibali 2013). In the field of developmental psychology, Church and Goldin-Meadow (1986) investigate (mis-)matching between gesture and speech according to the stages of cognitive development of children by conducting the experiments using a conservation task. Their study reveals that the number of gesture-plus-speech explanations slightly decreases as children get older, and children's use of discordant explanations (gesture expresses different information from that contained in speech) indexes inconsistency in their knowledge (judgement and explanations). Alibali and Kita (2010) also investigated gestures of children in conversation tasks and recognised the use of gestures when describing perceptual information. Alibali and Goldin-Meadow (1993), on the other hand, analysed children's use of gestures when solving math problems and found that children who used gestures learned better than the ones who did not. In a more recent study, Alibali and Nathan (2012: 277) examined gestures of learners and teachers in math classrooms, identifying three gestures used in the context:

1. pointing gestures reflect the grounding of cognition in the physical environment,
2. representational (i.e., iconic and metaphoric) gestures manifest mental simulations of action and perception, and
3. some metaphoric gestures reflect body-based conceptual metaphors.

These findings significantly benefit the current study. However, there are two aspects which are missing in the existing studies: (1) gestures tend to be analysed focusing only on speakers by collecting gestures in relatively short isolated instances, i.e. analysing a speaker's pointing gesture which reflects his/her cognition of the physical environment without considering a subsequent response from a listener, and (2) gestures used by learners of a second or additional language are not targeted in most studies. There is a limited number of studies which have examined gestures in language learners (see a review of those studies in the following section), but the collaborative use of gestures between interlocutors in relation to their proficiency levels is rarely focused on.

2.2 Gestures in repair sequences

Schegloff, Jefferson and Sacks (1977: 364–365) identify four types of repair: self-initiated self-repair (SISR), other-initiated self-repair (OISR), self-initiated other-repair (SIOR), and other-initiated other-repair (OIOR). Applying the notion of repair and the four categories, several studies address the question "how do interlocutors use gesture (gesticulation) when some problems, in other words 'trouble sources' (Schegloff 2007), occur in the ongoing conversation?" M. Goodwin and C. Goodwin (1986) analyse gaze, hand gestures and facial expressions in "the activity of searching for a word", which can be regarded as a repair, describing the coordination of these modes as shown in Figure 1. Speaker A is searching for a more precise word to describe the place than the word *bunks*. A vertical line above the words indicates A's gaze to B, which first appears the moment A says *in the*, and the second one starts during the pause (the conventional symbols '-' and '+' indicate the length of the pause in the example), continuing while A is uttering *in the room*. The underlined word is uttered with emphasis.

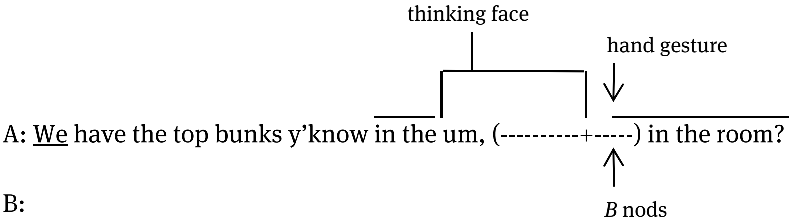

thinking face
hand gesture
A: <u>We</u> have the top bunks y'know in the um, (----------+-----) in the room?
B:
B nods

Fig. 1: Repair example adapted from M. Goodwin & C. Goodwin (1986: 31).

The example transcription depicts A's facial expression 'thinking face', which is co-produced with *um*, initiating the pause. A hand gesture was produced during the pause before the utterance *in the room*. B's head nods can be interpreted as an acknowledgement token that expresses B's adequate understanding of what A is trying to say although it is used before A provides the alternative phrase *in the room*. A's hand gesture, which is used during the pause with turning to B, is also taken as a meaningful action, which can be interpreted as a request for B's aid in the search. C. Goodwin (1981: 142) also identified the use of a lengthening sound to initiate a SISR: a speaker tends to produce prolongation of the sound of the last word until regaining the recipient's gaze, which can be "a repair initiation signalling, and preparing for, the upcoming restart".

Gestures in repair sequences in a second language (L2) setting are examined in Olsher (2004). His study investigates the use of hand gestures in L2 talk among Japanese learners of English in a classroom, focusing on 'embodied completion' which refers to "ceasing to talk and completing the action that had been initiated by the partial turn through gesture or embodied display" (Olsher 2004: 221). Olsher (2004: 243) regards the use of embodied completion as a demonstration of language learners' skills to successfully achieve mutual understanding and develop sequences of talk in action rather than a "limitation" or "failure". However, the relationship between the use of embodied completion and learners' language ability in L2 is not fully explored in his study.

Another study of gestures in L2 (English) conversations in relation to repair is reported in Seo and Koshik (2010). They examined gestures which were used to initiate repair in tutorials between a native speaker tutor and a non-native speaker tutee (a native speaker of Korean). In their study, two distinct gestures, a head poke and a sharp head tilt/turn, are recognised as repair initiators, which sometimes accompany verbal repair initiators, such as *They're what?* (Seo & Koshik 2010: 2222). Their analysis identified that this practice is shared between both native tutors and non-native tutees.

Hosoda (2006) also focuses on the use of repair in Japanese conversations between native (Japanese as a first language, JL1) and non-native speakers of Japanese (Japanese as a second language, JL2) in casual settings in relation to their (dis-)orientation to differential language expertise. In her JL1-JL2 conversation data, the JL2 speakers have an "apparent problem of producing a single lexical item" (Hosoda 2006: 37), which is often repaired by the JL1 speakers. Hosoda (2006: 38) also introduces 'the repair sequence format' (RSF) with repair observed in her data:

- Turn 1 (T1) Talk that contains a reparable item
- Turn 2 (T2) Other-initiation (OI) of repair
- Turn 3 (T3) Attempt at self-repair (SR)
- Turn 4 (T4) Other-repair (OR)
- Turn 5 (T5) Acceptance of OR in the form of repetition
- Turn 6 (T6) Return to main sequential action

Hosoda (2006) found that the explicit other-initiation of repair, such as the Japanese phrase *shiranai* ("I don't know [the word]"), is distinct in the JL1-JL2 conversations, which is rarely observed in L1 conversations. The JL1 speakers orient to their language 'expert' role by taking up the JL2 speakers' repair-initiation, while the JL2 speakers take their 'novice' role to accept the repair and repeat the suggested word by the JL1 speakers. However, a multimodal analysis of the use of hand gestures is limited in her study. Taleghani-Nikazm (2015: 86) conducted a single case analysis of a natural group conversation by L2 learners of German, finding that pointing gestures are used when the speakers offer missing words for others and a particular gesture is observed, when a speaker signals their engagement in the activity of word search and simultaneously requests 'recipient collaboration' with other-repair (Taleghani-Nikazm 2015: 98). Her study includes the aspects of multimodality, but it is a single case study and the relationship between the learners' use of gestures and their proficiency is not addressed.

To fill the gap, the current study attempts to establish a methodology to analyse how Japanese learners of English use gestures in a longer interaction sequence in a dyad interaction, focusing on practices of repair, comparing English conversations of an advanced-level pair and a basic-level pair using CGCA.

3 Research data and methods

For the analyses of the current study, two five-minute-long pair conversations in English as a L2 were recorded at a university in Japan, one of which is carried out by a pair of advanced learners (B1 in CEFR) and the other by basic-level learners (A2). All four participants were second-year Japanese students in engineering and science departments. These two pairs were enrolled in different classes of the English communication programme at the university according to their proficiency levels. The advanced-level pair consists of Sota (male) and Kazu (male) and the basic-level pair Yuri (female) and Koji (male). These names are pseudonyms. The participants in each pair had known each other for the period of their English class (i.e. four months at the time of recording), although they met only in the English class twice a week. Both pairs were asked to talk about their plan for the coming summer holidays in the conversations, which were conducted as practices for the oral test of the English programme. They were given five minutes preparation time before they started the conversations.

The two data sets were compared in terms of word count, speaker turns, and speaking time in order to gain an overview of the interaction in each pair. I define 'speaker turns' to consist of more than or equal to three words. The speaking time lengths are calculated for the quantitative analysis using the time-aligned transcripts. Frequency and functions of gestures and repair strategies were also analysed both in quantitative and qualitative approaches. using a method that I call 'conversational gesture corpus analysis' (CGCA), which was developed on the basis of the time-aligned multimodal corpus (for a detailed description of this type of corpus, see Tsuchiya 2013). CGCA consists of two steps:

1. The transcribed data of the participants' utterances and gestures are stored in a multimodal corpus for the quantitative analysis (i.e. word count and speaking time length). The time-aligned scripts visualise participants' utterances and gestures on a timeline (in seconds) to capture longer sequences in interaction with multimodal features, which enables researchers to identify interesting areas for micro-level investigation.
2. The targeted instances identified in the previous stage, i.e. hand gestures in repair sequences in this study, are extracted and analysed with the temporality and functions of gestures in detail with Praat and the video images, applying a conversation analytic approach.

The functions of gestures were annotated by referring to McNeill (1992) and Goldin-Meadow's (1999) classification. Although there are some instances for which it is difficult to distinguish whether they are 'beat' or 'metaphoric' gestures in the current data, I decided to label the gestures apparently rhythmic without any representational functions as 'beat gestures', and those representational, or both rhythmic and representational (the ambiguous cases), as 'metaphoric gestures'.

Each participant's utterances and gestures were transcribed and time-stamped using the software package Transana version 2.12 (Fassnacht & Woods 2002) and combined as time-aligned transcripts on the second time scale as shown in Table 1. The first column in Table 1 is the timeline in seconds (the last two digits are the seconds). Sota's gestures and utterances appear in the second and third columns. In the third and fourth columns, Kazu's gestures and utterances are aligned (HN indicates his use of headnods although the primary focus is placed on hand gestures in this chapter. MG indicates a metaphoric gesture. I added the words which are accompanied by the gestures in brackets). In the last column, I annotated instances of 'repair', SISR (self-initiated self-repair), for example.

Tab. 1: Time-aligned transcript of Extract 1; the beginning of the advanced-level conversation.

Time(sec)	Sota Gesture	Sota Transcription	Kazu Gesture	Kazu Transcription	Repair
00: 00: 17	HN	Mhm.		Er but er I could= can't study abroad in next seme= next semester(.) <$E> laugh </$E> Er: because I fail the=	
00: 00: 18					
00: 00: 19					
00: 00: 20			MG (fail)		
00: 00: 21	HN				
00: 00: 22					
00: 00: 23	HN				
00: 00: 24					
00: 00: 25					
00: 00: 26	HN	|Yeah.	MG (interview test)	|interview tests. Erm so er I have many choices now er such as er entering another university+	
00: 00: 27					
00: 00: 28					
00: 00: 29					
00: 00: 30					
00: 00: 31					
00: 00: 32	HN				
00: 00: 33					
00: 00: 34					
00: 00: 35					
00: 00: 36					
00: 00: 37	HN	Mhm.		+or studying abroad in (.) a new (.) |term.	
00: 00: 38					
00: 00: 39					
00: 00: 40					
00: 00: 41			MG (new term)		SISR
00: 00: 42					
00: 00: 43		|New term?			

When hand gestures are analysed here, time spaces between movements are counted based on the methodology established in Tsuchiya (2013). If there are several hand gestures within one second, they are counted as a hand gesture since the timeline in seconds was applied to the time-aligned transcripts. If a hand gesture is continuously used for more than a second, this gesture is divided into two gestures according to the time scale. By doing so, the time-aligned scripts visualise the multiple participants' utterances and gestures in a timeline and capture a longer sequence in interaction (see Table 1). This enables researchers to identify interesting areas for further investigation with a micro-level analysis.

Annotations of repair strategies were added manually to the transcripts. The timing of a gesture in a repair sequence was also visualised using a phonetic analysis tool, Praat version 5.3.83 (Boersma & Weenink 1992–2014), for the micro-analysis of the temporality of linguistic and gestural features in repair. Thus, the combination of the global pattern analysis with the time-aligned corpus and the micro-level analysis with a conversation analytic approach is the novel aspect of the present approach.

4 Results

This section reports the results of the comparative analysis between the advanced-level pair conversation and the basic-level counterpart from three aspects: (1) their interactional patterns (numbers of turns and words), (2) the number of hand gestures in repair sequences, and (3) the timing of hand gestures in repair.

4.1 Numbers of turns and words in the conversations

Numbers of turns, words, and speaking time of each participant in the two pairs are summarised in Table 2. Sota and Kazu formed the advanced-level pair. The total word count in the five-minute conversation is 486, which is slightly greater than that of the basic-level pair (427 words). Sota in the advanced-level pair took six turns and spoke for two minutes 44 seconds in total. His average speaking time per turn is 27 seconds, which is about 10 seconds longer than that of his conversation partner, Kazu, who had more turns than Sota (8 compared to 6). The numbers of turns of the two participants in a pair are not equal since I counted an utterance with more than three words as a turn and excluded response tokens, such as *yeah* and *mhm*, since, if I include them, it affects the average length of turns.

Tab. 2: Numbers of turns, words and speaking time of the advanced- and basic-level pairs.

		Speaking time	Speaker turn	Word count	Average speaking time per turn
Advanced	Sota	0:02:44	6	279	0:00:27
	Kazu	0:02:06	8	207	0:00:16
	Pause	0:00:10	-	-	-
	Total	0:05:00	14	486	-
Basic	Yuri	0:02:29	20	239	0:00:07
	Koji	0:01:52	9	188	0:00:12
	Pause	0:00:39	-	-	-
	Total	0:05:00	39	427	-

The conversation of the basic-level pair includes more pauses (39 seconds in total) than the advanced pair (10 seconds). Unequal participation between the participants in the basic-level pair is more obvious than for the advanced pair. Yuri dominates the interaction, taking 20 turns and speaking for two minutes and 29 seconds in total, while Koji takes turns only nine times and speaks for less than two minutes in total. The average speaking time lengths of both participants in the basic-level group are shorter than those of the advanced-level pair.

The results from the quantitative analysis using the CGCA method indicate that there are differences in the two pairs in their turn-taking patterns: the advanced-level pair has fewer and longer speaking turns than the basic-level pair. The two participants in the advanced-level pair show a comparative amount of contribution, whereas asymmetrical participation and more pauses are recognised in the conversation of the basic-level pair.

4.2 Numbers of hand gestures in repair sequences

For comparison, occurrences of hand gestures by each participant are counted and summarised in Table 3, classified into the four functions. The total number of the instances of gestures in the advanced-level pair conversation is 74, which is more than that of the basic-level pair (55 instances in total). Sota in the advanced-level pair uses gestures about six times more than Kazu. Most of Kazu and

Sota's gestures are classified as metaphoric gesture. However, Sota also uses iconic gestures and deictic gestures.

Tab. 3: Frequency of gestures per function.

		IG	MG	BG	DG	Total
Advanced	Sota	8	54	0	1	63
	Kazu	0	9	2	0	11
	Total	8	63	2	1	74
Basic	Yuri	18	8	0	5	31
	Koji	7	9	0	8	24
	Total	25	17	0	13	55

Note: IG = iconic gesture, MG = metaphoric gesture, BG= beat gesture, DG = deictic gesture

Fewer occurrences of gestures were observed in the conversation of the basic-level pair: Yuri uses gestures 31 times in total and Koji 24 times. The advanced- and basic-level learners seemed to use gestures for different functions. Metaphoric gestures are used most frequently (63 times) in the conversation of the advanced pair, while in the basic-level pair, iconic gestures (25) are used more than metaphoric gestures (17). Another difference is that the use of deictic gestures is limited in the advanced-level pair (only once), whereas the basic-level pair conversation includes 13 occurrences of deictic gestures. This point will be investigated further in Section 4.2 in relation to repair strategies.

The number of the occurrences of repairs in the two pairs is shown in Table 4. To analyse the relationship between the use of the four types of repair strategies and functions of co-occurring gestures, the number of the gestures and their functions is indicated in brackets (see Table 4). In most cases, the participants repair verbally by speech with support of visual hand gestures. There is, however, one instance where a gesture is produced during a silent pause without any accompanying speech and at the same time initiates an other-repair[1].

[1] On this point, I would like to thank members of the ICAME35 audience for their valuable feedback.

Tab. 4: Frequency of repair instances with gestures.

		SISR	SIOR	OISR	OIOR	Total
Advanced	Sota	7 (7MG)	0	0	0	7
	Kazu	2 (2MG)	0	0	0	2
	Total	9	0	0	0	9
Basic	Yuri	5 (3IG, 2MG)	0	1 (w/o HG)	0	6
	Koji	7 (3DG, 2MG, 2IG)	2 (1IG, 1DG)	0	0	9
	Total	12	2	1	0	15
Total		21	2	1	0	24

Note: Acronyms are used to express the four repair strategies (SISR = self-initiated self-repair, SIOR = self-initiate other-repair, OISR = other-initiated self-repair, and OIOR = other-initiated other-repair) and the four functions of gestures (MG = metaphoric gesture, IG = iconic gesture, BG= beat gesture, DG = deictic gesture). The acronym, w/o HG, indicates an instance of repair without any hand gesture.

In the advanced-level pair, nine instances of SISR with metaphoric gestures were observed. Sota initiates and self-repairs seven times, which is three times more than that of Kazu. All the trouble sources were single lexical items that they had some difficulty to produce in English. The conversation of the basic-level pair includes more occurrences of SISR than the advanced-pair conversation: Koji self-repaired seven times and Yuri five times. The other difference is the variety of the co-produced gestures in the basic-level pair: iconic gestures (five times) are more frequently observed in repair sequences than metaphoric gestures (four times) and deictic gestures (three times). Koji, the trouble source speaker, also initiates other-repair twice, once with an iconic gesture and the other time with a deictic gesture, both of which were repaired by Yuri (SIOR). Only one occurrence of OISR without hand gestures was observed in the basic-level pair, which was initiated by Yuri to repair a repairable that she recognised in Koji's previous utterance. This is the only occurrence of the participants in the basic pair repairing the content rather than single-word lexical items.

4.3 Timing of hand gestures in repair sequences

The temporal relationship between speech and accompanying hand gestures in the repair sequences will be investigated further through the qualitative analysis in this sub-section. As examined in Section 4.2, metaphoric gestures were observed in the course of SISR in the advanced-level pair conversation. Extract 1 is such an example where hand gestures are co-produced with speech as a repair initiation. Extract 1 is the beginning of the conversation, and before this part, Sota first asks Kazu about his plans for the summer holidays. Then Kazu starts talking about his plan – to study English in order to join a study abroad programme – providing some reasons. From 00:00:17 to 00:00:37, Kazu says that he planned to study abroad from the next semester but he failed the interview test, and he mentions some other alternatives he could take. Sota inserts continuer response tokens twice: *yeah* at 00:00:26 and *mhm* at 00:00:37. At 00:00:37, Kazu produces a metaphoric gesture with the word *new* before uttering the missing word *term*, which is followed by a request for confirmation by Sota in line 00:00:43: *new term?*

Extract 1: Kazu's SISR with metaphoric gesture in the advanced-level pair.[2]

00: 00: 17	Kazu		Er but er I could= can't study abroad in next seme= next semester (.) <$E> laugh </$E> Er: because I fail the+
00: 00: 26	Sota		|Yeah.
00: 00: 26	Kazu		|interview tests. Erm so er I have many choices now er such as er entering another university+
00: 00: 37	Sota		Mhm.
00: 00: 37	Kazu	→	+or studying abroad in (.) er new (.) |term. <$E> MG by Kazu </$E>
00: 00: 43	Sota		|New term?

A detailed description of the organisation of Kazu's gestures and the associated words in Extract 1 is provided in Figure 2.1. Kazu starts the gestures during the short pause before *new*, which is indicated as 'onset' in Figure 2.1, referring to the three phases of a gesture defined by Kendon (1980) and McNeill (1992) (see Section 2.1). As shown in Figure 2.2, Kazu lowers his right hand from his neck, where he rests the hand, towards the desk in front of him as he looks down, and slides his hand quickly from his left to right with his palm open twice while he utters *new*, which is prolonged slightly, similarly to "the activity of searching for a word" that M. Goodwin and C. Goodwin (1986: 56) find (see Section 2.2). Then he

[2] The plus symbol + indicates a continuous sentence and the equal symbol = signals an unfinished sentence. <$G?> indicates inaudible sounds and | indicates overlap between a previous speaker and a following speaker. <$E>…</$E> shows extralinguistic information including laughter, cough and notes, and <$H>…</$H> appears where the accuracy of the transcription is uncertain. (2.0) indicates an interval between utterances (2 seconds in this case) and (.) indicates a very short untimed pause. An arrow (→) at the beginning of a line indicates that the line is particularly important for CGCA.

stops the movement of the stroke and rests his right hand with his palm open on the desk and, after another short pause, he utters *term*, looking at Sota at the same time. I take this as an action of searching for a word and SISR. When Kazu cannot say the word *term* immediately, he takes a pause, produces the hesitation marker *er*, and lengthens the sound of the precedent word *new* with the sliding hand gesture, which seems to function as a repair initiation. After the hand gesture, he utters the missing word *term*.

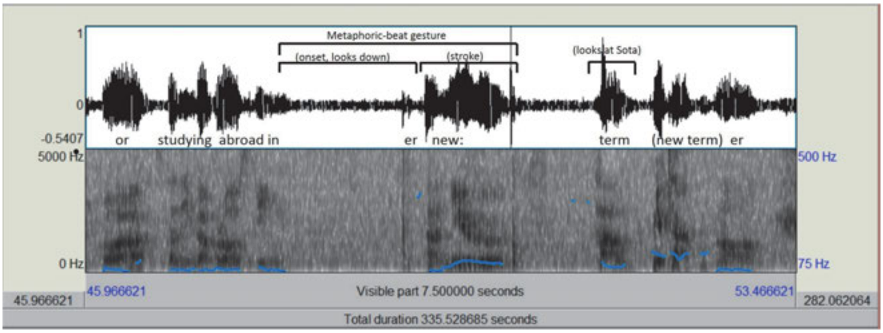


Fig. 2.1: Timing of Kazu's metaphoric gesture.

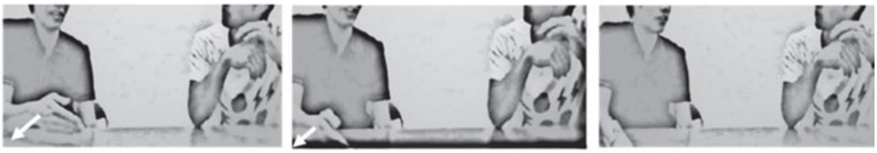


Fig. 2.2: Kazu's metaphoric gesture.

Extract 2 is a second example of metaphoric gestures in particular observed in the advanced-level pair. Here, responding to Kazu's question, Sota starts talking about his plans for the summer holidays at 00: 01: 22. Then he explains why he has to study hard, referring to his future career and his intention to change his course. At 00: 01: 50, while he is listening to Sota, Kazu produces a short and shallow head nod, which functions as a continuer response token.

Extract 2: Sota's SISR with metaphoric gesture in the advanced-level pair.

00: 01: 21	Kazu		And you?
00: 01: 22	Sota		Erm: so: I will study hard too so: er: so because= this is because so: actually I want to be a <$E> occupation </$E> in the future so er: so: ne= next year so I want to= I want to change my course from er: <$E> course name </$E> er: <$E> course name </$E> ah? <$E> course name </$E> to erm this <$E> a course name </$E> course. So: erm so:+
00: 01: 50	Kazu		<$E> Kazu gives a short shallow headnod </$E>
00: 01: 54	Sota	→	+the er: so er exam? exam? or er the requirement+ <$E> MG by Sota </$E>
00: 02: 02	Kazu		<$E> Kazu gives a short shallow headnod </$E>
00: 02: 03	Sota		+is the= er so taking= taking= er: TOEFL iBT scores over 48. So er I= so I= I studied= erm so for iBT.

At 00: 01: 54, Sota tries to explain the requirements he needs to meet in order to transfer from the current course to the course he wishes to join. Sota, however, has some difficulty in uttering the word *requirement* at first, and initiates a repair sequence by uttering the alternative word *exam* twice with a rising tone, inviting a repair from Kazu. The practice of the repetition of the marked word is also reported in Hosoda (2006) in her Japanese conversation data. However, Kazu keeps looking away and does not co-operate in Sota's activity of searching for a word, so that Sota then self-repairs, uttering the missing word *requirement*. This is another example of an activity of searching for a word and SISR.

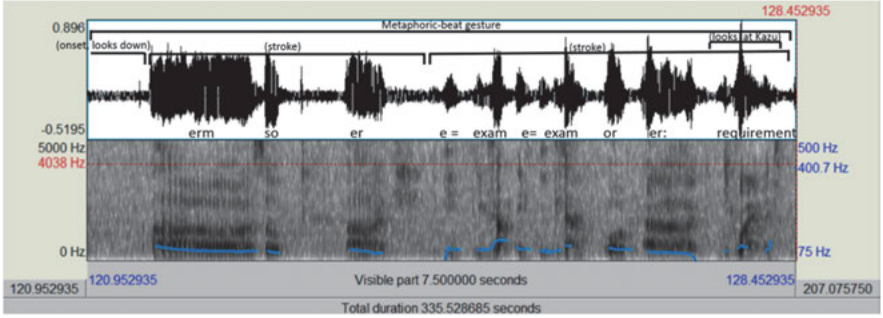


Fig. 3.1: Timing of Sota's metaphoric gesture.

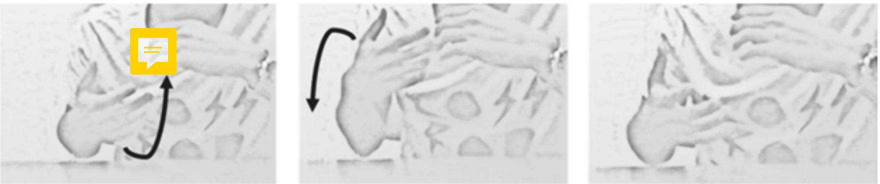


Fig. 3.2: Sota's metaphoric gesture.

Before the repair sequence, Sota has joined his hands in front of him, lifting them from the desk at a lower position (Figures 3.1 and 3.2). At 00: 01: 54, he looks down and lifts both of his hands slightly with his palms open and circulates them several times from the bottom to the top one after another in coordination as if he was drawing a small circle with both hands in the air. This hand gesture continues while he is producing hesitations, such as a prolonged *erm* and *so* before he utters the word *exam*. When he utters *e= exam* twice with the hesitation *e=*, he stops moving his left hand around his chest and rotates only his right hand twice, synchronising with the sounds of the two instances of the word *e= xam*. He then stops the movement of his right hand shortly with his palm facing upwards when he pronounces the nucleus of the word *exam* (/igzǽm/) with emphasis, simultaneously looking at Kazu briefly. Then he produces the same circular hand gestures now with both hands and utters the missing word *requirement*, turning to Kazu, which is followed by Kazu's short and shallow head nod to express his acknowledgement. During the five-minute conversation, Sota frequently uses hand gestures, especially the circulation gesture with both hands.

As described above, the participants in the conversation of the advanced pair produce metaphoric gestures in the sequences of SISR. In some instances, Sota attempts to initiate other-repair. In other words, Sota requests language support from Kazu, which is similar to the request of recipient collaboration (Taleghani-Nikazm 2015, see Section 3.2), but it is not taken up by Kazu, who seems to avoid the 'language expert' role in the interaction.

In the conversation of the basic-level pair, iconic gestures and deictic gestures are also produced in the course of repair activities. Extract 3 includes an occurrence of Yuri's iconic gestures in SISR. At 00: 02: 52, Yuri started talking about her summer holiday plans, saying *yeah this summer vacation I will go hometown in <$E> city name </$E>*, which is followed by Koji's continuer response token *un un* at 00: 03: 01 with head nods.

Extract 3: SISR with iconic gestures in the basic-level pair.

00: 02: 52 Yuri Yeah this summer vacation I will go hometown in
<$E> city name </$E>.
00: 03: 01 Koji Un un. <$E> head nods </$E>
00: 03: 02 Yuri → So I and my mother go= we will go shopping. Maybe.
<$E> DG and IG by Yuri </$E>

When Yuri says *I and my mother*, she produces two deictic gestures, first pointing herself on the chest at the timing when she says *I* and the front with her right hand with palm open, synchronously uttering the words *my mother*. Then she first says the verb *go* without *will*, but recognises the error and self-repairs immediately. At the same time, she uses iconic gestures, which are depicted in Figures 4.1 and 4.2.

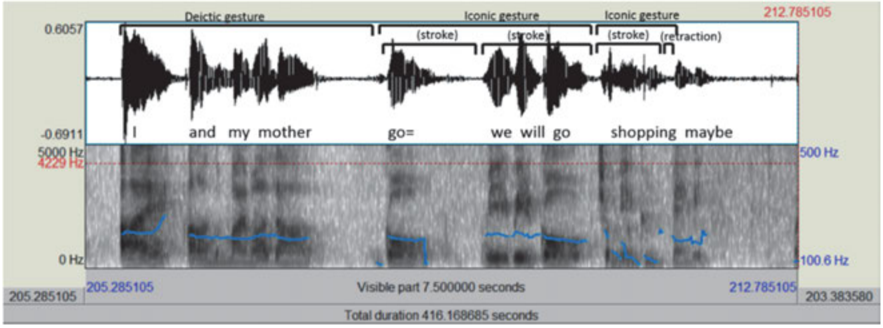


Fig. 4.1: The timing of Yuri's iconic gesture.

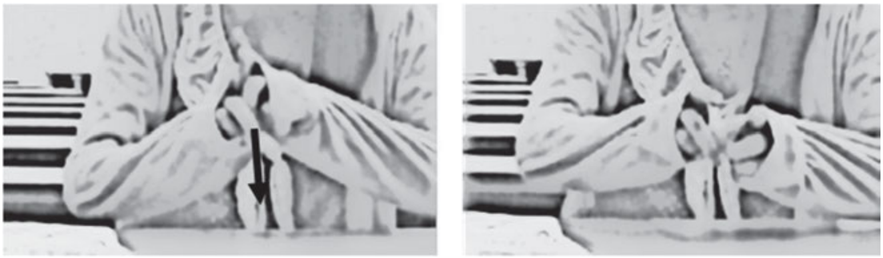


Fig. 4.2: Yuri's iconic gesture.

When Yuri says *go=*, she joins her hands at a lower position with her index fingers up and moves both hands forward to describe the action of her and her mother, expressed by her two index fingers, going to the shopping site together. This iconic gesture is an O-VPT gesture since the gesture describes the movement from an observer's view point (McNeill 1992). Yuri repeats the gesture when she self-repairs. Yuri also produces a different iconic gesture when she utters the word *shopping*: she swings her arms slightly several times as if she was walking, which is a C-VPT iconic gesture. As described here, a sequential use of iconic and deictic hand gestures was observed in her speaker turns (also see Table 3 in Section 4.2). Thus, the basic-level learners use more deictic gestures to refer to themselves or others in the conversation.

Koji also produces hand gestures frequently in the basic-level pair conversation. An instance of Koji's deictic gesture is observed in Extract 4. Before this extract, Koji talks about his plan to go camping in summer with his friends who are members of a university club he joined. At 00: 04: 10, Yuri asks whether he is going to have a barbecue at the camp.

Extract 4: SISR with deictic gesture in the basic-level pair.

00: 04: 10	Yuri		Will (.) you (.) do barbecue?
00: 04: 15	Koji		Barbecue (.) oh I= (.) | ee:
00: 04: 18	Yuri		| Maybe?
00: 04: 19			Pause (3.0)
00: 04: 20	Koji	→	This cir= I belong circle <$E> DG by Koji </$E> is= ee:
00: 04: 25			Pause (3.0)
00: 04: 28	Koji	→	=Did <$E> MG by Koji </$E> barbecue in (.) summer camp ee: *MAITOSHI*. <$E> In Japanese (every year) </$E> <$E> laugh </$E>
00: 04: 33	Pause		(1.0)
00: 04: 34	Yuri		Every year.
00: 04: 35	Koji		Every year.

Koji repeats the word *barbecue* with a soft falling tone for confirmation, and after a short pause, he utters *oh*, which is a change-of-state token (Heritage 1984), and tries to start the next turn with *I=*. After the following short pause, at 00: 04: 18, Yuri utters *maybe*, which is overlapped with Koji's hesitation *ee:*. Both of them pause for a second to negotiate the next speaker turn due to the overlap. At 00: 04: 20, Koji takes the floor, initiating a self-repair with a deictic gesture (see Figures 5.1 and 5.2).

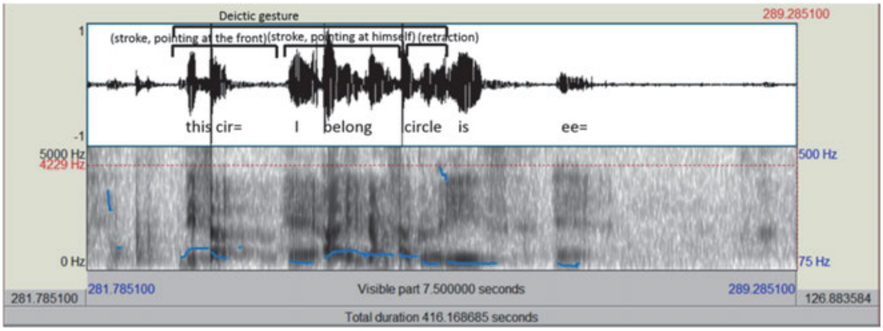


Fig. 5.1: The timing of Koji's deictic gesture.

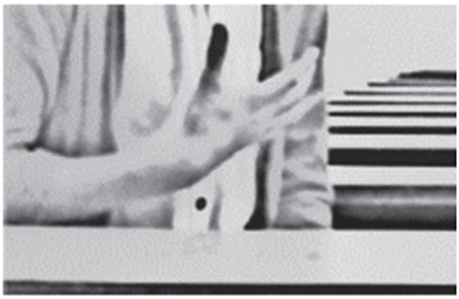


Fig. 5.2: Koji's deictic gesture.

When he says, *This cir=*, he slightly lifts his right hand and lowers the hand toward the desk pointing to the front with his index finger. He then moves his right hand toward his chest with his palm open, providing a self-repair *I belong circle is=*, by which he means "the circle [club] I belong to".

After the three seconds pause with his 'thinking face', which is the term in M. Goodwin and C. Goodwin (1986) (see Section 2.2), at 00: 04: 28, Koji self-repairs again and corrects the verb *is*, which he has wrongly chosen in his previous utterance, uttering *did barbeque* with a metaphoric gesture, moving his hands from his front to the left. Another repair strategy, the use of his first language (L1, Japanese), is also observed, which initiates an other-repair. He utters *Maitoshi* (every year) in Japanese at 00: 04: 28 with laughter, which fills a second pause. Responding to Koji's request for other-repair, Yuri offers the missing word *every year* at 00: 04: 34, which is followed by Koji's repetition to accept Yuri's repair at 00: 04: 35. The use of their L1 (Japanese) was limited in both conversations since they were encouraged to avoid using L1 during the practice and in the oral exam in the

English course. Therefore, the use of the Japanese word *Maitoshi* by Koji in Extract 4 was marked with his laughter and the following silence. Here, Yuri takes up her ‘language expert’ role through the interaction, while Koji acts the part of a ‘dependent language learner’.

All of the examples of the use of gestures in repair sequences described above are accompanied by speech. However, there is one occasion where a gesture without any associated words is used as an other-repair initiator in the basic-level pair conversation, which is shown in Extract 5.

Extract 5: SIOR with iconic gesture during a pause in the basic-level pair.

<table>
<tr><td>00: 01: 08</td><td>Yuri</td><td></td><td>Can I listen this podcast my computer? <$E> IG by Yuri </$E></td></tr>
<tr><td>00: 01: 14</td><td>Koji</td><td></td><td>U:n. this podcast title is <i>MINNA NO YUME NEWS</i></td></tr>
<tr><td></td><td></td><td></td><td><$E> In Japanese (News of Everyone’s Dream)</$E></td></tr>
<tr><td></td><td></td><td></td><td><$E> MG by Koji </$E></td></tr>
<tr><td>00: 01: 19</td><td>Yuri</td><td></td><td>Okay okay. <$E> Head nods by Yuri </$E> erm (.)</td></tr>
<tr><td>00: 01: 20</td><td>Pause</td><td>→</td><td>(3.0) <$E> IG by Koji </$E></td></tr>
<tr><td>00: 01: 23</td><td>Yuri</td><td></td><td>Okay okay | I check I check. <$E> laugh and IG </$E></td></tr>
<tr><td>00: 01: 24</td><td>Koji</td><td></td><td>|<$H> rep= </$H> <$E> laugh </$E> check u:n.</td></tr>
</table>

At the beginning of the conversation, Koji talks about a podcast recording that he was involved in as an editor recently. At 00: 01: 08, Yuri asks whether she can listen to the podcast on her computer, synchronously providing some iconic gestures: pointing her ears with both hands when she says *listening* and using a pantomime (typing on a computer keyboard with both hands) when she utters *computer*. Koji answers with a hesitation markers *u:n*, and at the same time moves his right hand onto his left hand which rests on the desk in front of him. Yuri inserts an acknowledgement response token *okay okay* in a soft voice with head nods at 00: 01:19, which is followed by a three-second pause. At the beginning of the pause, Koji wears a ‘thinking face’ while searching for a word, and then smiles, simultaneously producing the same iconic hand gesture Yuri just used (typing) without words, and turns to Yuri while holding the gesture, which is a C-VPT iconic gestures and functions as an other-repair initiator (see Figures 6.1 and 6.2).

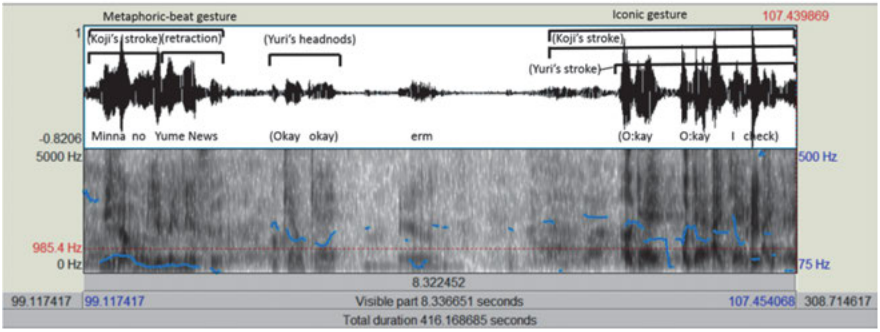


Fig. 6.1: Timing of Koji's iconic gesture in a pause.

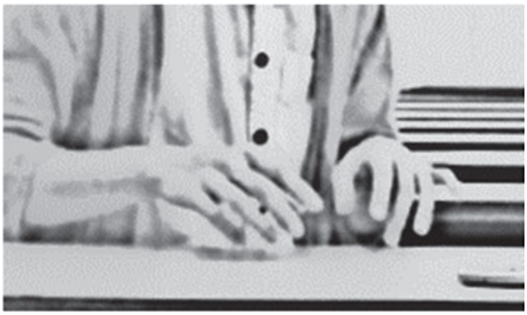


Fig. 6.2: Koji's iconic gesture in a pause.

At 00: 01: 23, Yuri responds *okay okay* and offers the missing word *I check* and repeats this phrase with laughter, providing the same iconic hand gesture (typing) again to accommodate Koji's gesture. Koji utters *rep=* at 00: 01: 24, which forms part of the word *report*, but he does not finish the word since it overlaps with Yuri's utterance *I check*. He laughs with Yuri and repeats the word *check* provided by her to accept the repair. He does not use his L1 in Extract 5. However, the use of a hand gesture without verbal expressions is also marked with laughter by both participants since they are supposed to express themselves verbally in the context of the oral examination and the practice.

Both participants in the basic-level pair conversation produce a variety of hand gestures in repair sequences and the use of SIOR is also observed in their interaction. Koji tends to initiate other-repair and Yuri takes up the 'language expert' role, responding to his request.

5 Discussion

The discussion consists of two parts: (1) a review of the research methodology of CGCA, its practical procedures, feasibility and limitations, and (2) a discussion of the preliminary results of the analysis from the perspective of learners' social roles in the interaction.

5.1 Conversational Gesture Corpus Analysis

There are two steps in the process of CGCA:

1. Transcription of the data and capturing of transcriptions in a corpus with time-aligned visualisation to identify examples for detailed analysis.
2. Qualitative analysis of selected examples with the help of Praat.

Tab. 5: Methodologies for a multimodal analysis.

	Quantitative Analysis	**Qualitative Analysis**
Multimodal Corpus Linguistics	The occurrences of hand gestures and head movements in interaction are quantitatively analysed in relation to spoken language, using an application which aligns multimodal elements in interaction.	A qualitative analysis of sequences in interaction may be limited or excluded from the focus.
Conversation Analysis	A quantitative analysis, such as the numbers of hand gestures in interaction, is excluded.	Detailed descriptions of sequences and structures of interaction are provided to reveal a social and institutional order.
Conversational Gesture Corpus Analysis	The occurrences of hand gestures in interaction are quantitatively analysed in relation to spoken language, using a time-aligned transcript. The types of hand gestures and repair strategies are also annotated in the transcript.	Interesting phenomena in interaction are highlighted and analysed further with a conversation analytic approach.

Thus, the CGCA enables researchers to find potentially interesting phenomena through a quantitative MCL method, which will be examined further using a qualitative CA approach. The method compensates for the weakness of MCL and CA if they are used alone. Table 5 illustrates the methodological comparison.

An area to be improved in future research is the taxonomy of gesture. In this study, the typology in Goldin-Meadow and McNeill (iconic, metaphoric and deictic gesture, C-VPT/O-VPT) were annotated for the preliminary research. Considering semantic features of gestures, i.e. 'action', 'entity', or 'shape' (Bergmann & Kopp 2006: 92–93), and interactional gestures, i.e. gestures that express objects, perceptual information and action (Alibali & Kita 2010)[3], could benefit further analysis.

5.2 Learner repairs and their social roles

The small-scale case study reported in this chapter reveals that the Japanese learners of both pairs employed hand gestures in almost all the instances of repair in the interactions. There are, however, some differences in the way they treat the trouble sources and their use of hand gestures when they repair. The results from the preliminary case study with the CGCA method are summarised below:

1. In terms of turn-taking structure, fewer but longer turns were observed in the advanced-pair conversation compared with the basic-level pair.
2. The numbers of gestures each participant used vary. Metaphoric gestures were produced more frequently in the advanced-level pair conversation, while the basic learners used iconic gestures most in addition to metaphoric gestures and deictic gestures.
3. The conversation of the basic-level pair includes more occurrences of repair than that of the advanced-level pair. In terms of the relationship between the repair practices and the use of hand gestures, all the instances of repair in the advanced-level pair conversation were categorised into SISR with metaphoric gestures, whereas the basic-level pair conversation included more occurrences of SISR with different functions of accompanying gestures and also two instances of SIOR. Sota in the advanced-level pair conversation used SISR more frequently than his conversation partner Kazu. Only Koji in the

[3] I thank the reviewers for their insightful comments and the useful references they suggested.

basic-level pair initiated other-repairs and his conversation partner Yuri repaired, responding to Koji's repair-initiation.

The relationship between learners' use of repair strategies with accompanying gestures and their representations of social identities were identified through qualitative micro-analysis. Sota in the advanced-level pair, for instance, tried to initiate other-repair in Extract 2, uttering the word *exam* twice in a rising tone with MG, which is not the exact word he was searching for at that moment. However, Kazu did not take up his repair-initiation and Sota self-repaired instead of waiting for Kazu's support. By doing so, Sota and Kazu seem to position themselves as 'independent leaners/language users' in an equal relationship. Koji in the basic-level pair, on the other hand, initiated other-repair twice, both of which were taken up by Yuri. Here, the participants oriented to their differential language expertise (Hosoda 2006): Koji seems to represent a 'dependent language learner' identity and Yuri an 'expert' identity.

These practices can be explained from a perspective of social interactional theory by referring to the concepts of 'interactional tension', which was introduced by Goffman (1961). The oral exam and the practice can be the context which involves interactional tension, or 'dysphoria', which refers to "a sensed discrepancy between the world that spontaneously becomes reality, and the one in which he is obliged to dwell" (Goffman, 1961: 40). Goffman (1961: 41) identifies two ways to maximise easing of the tension: (1) granting the character of the activity; and (2) determining the most effective allocation of internally-generated resources. In the case of the basic-level pair conversation, Koji is taking a learner role, which is one of the social roles available and accepted in the context of practising English speaking, where the participants are assumed as learners, as taken for granted by the performers and the audience. Accordingly, Koji enacted one of his available social roles of a novice or dependent learner by initiating the sequences of SIOR, which was co-constructed with Yuri's acceptance of the "relatively expert" role. This is, however, not the case in the advanced-level pair since Kazu refused to take up the "relatively expert" role which is requested by Sota through his attempt to initiate an other-repair.

6 Conclusion

This study contributes to the development of the methodology of CGCA, which integrates the features of MCL and CA. CGCA provides an overview of the distinctive features of interactions in terms of turn-taking patterns and the use of gestures through a quantitative analysis. A qualitative conversation analytic approach is also applied to conduct a micro-level analysis for more detailed descriptions of the practice highlighted in the quantitative analysis, such as the temporality of the use of hand gestures in repair sequences.

Applying the method of CGCA, a small-scale multimodal corpus was developed for the comparative analysis of the use of hand gestures in repair sequences in English as L2 conversations between the advanced-level Japanese learner pair and the basic-level counterpart for the preliminary study. Some differences were observed in the use of gestures between the two pairs, i.e. the advanced learners self-repaired with metaphoric gestures, while the basic learners other-repaired with iconic gestures. This seems to be related to the learners' representations of different social roles in interaction. Finer descriptions of language learners' use of gestures in relation to practices of repair can be obtained with a larger data set. With the methodology developed in this study, it is hoped that future research brings a better understanding of the use of hand gestures in learner-learner interaction and its relationship with language learners' communication skills and construction of social identities in talk-in-interaction.

References

Adolphs, S., & Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Abingdon, UK: Routledge.

Adolphs, S., Knight, D., & Carter, R. (2011). Capturing context for heterogeneous corpus analysis. International Journal of Corpus Linguistics, 16(3), 305–324.

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*(4), 468–523.

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture, 10*(1), 3–28.

Alibali, M. W., & Nathan, M., J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences, 21*(2), 247–286.

Baldry, A., & Thibault, P. (2006). Multimodal corpus linguistics. In G. Thompson & S. Hunston (Eds.), *System and Corpus: Exploring Connections* (pp. 164–183). London, UK: Equinox.

Beattie, G., Webster, K., & Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. *Journal of Language and Social Psychology, 29*(2), 194–213.

Bergmann, K., & Kopp, S. (2006). Verbal or visual? How information is distributed across speech and gesture in spatial dialog. In D. Schlangen & R. Fernandez (Eds.), *Proceedings* *of the 10th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 90–97). Potsdam, Germany: Universitätsverlag.

Boersma, P., & Weenink, D. (1992–2014). Praat. Version 5.3.83. Retrieved from www.praat.org (last accessed August 2018).

Butterworth, B., & Hadar, U. (1989). Gesture, speech and computational stages: A reply to McNeill. *Psychological Review, 96*(1), 168–174.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics, 37*(6), 871–887.

Church, B., R, & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*(1), 43–71.

Fassnacht, C., & Woods, D. (2002). Transana. Version 2.12 - Win. University of Wisconsin-Madison, Madison, WI.

Goffman, E. (1955). On face work: An analysis of ritual elements in social interaction. *Psychiatry, 18*(3), 213–231.

Goffman, E. (1961). *Encounters*. Harmondsworth, UK: Penguin Books.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*(11), 419–429.

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology, 64*, 257–283.

Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press.

Goodwin, C., & Heritage, J. (1990). Conversation analysis. *Annual Review of Anthropology, 19*, 282–307.

Goodwin, M., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica, 62*(1/2), 51–75.

Halliday, M. A. K. (1978). *Language as Social Semiotic*. London, UK: Arnold.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In M. J. Atkinson & J. Heritage (Eds.), *Structures of Social Actions: Studies in Conversation Analysis* (pp. 299–345). Cambridge, UK: Cambridge University Press.

Hosoda, Y. (2006). Repair and relevance of differential language expertise in second language conversations. *Applied Linguistics, 27*(1), 25–50.

Jewitt, C. (2009). Different approaches to multimodality. In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (pp. 28–39). London, UK: Routledge.

Jewitt, C., & Jones, K. (2008). Multimodal discourse analysis: The case of 'ability' in UK secondary school English. In V. Bhatia, K. J. Flowerdew, & R. H. Jones (Eds.), *Advances in Discourse Studies* (pp. 149–160). London, UK: Routledge.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In M. R. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207–227). The Hague, Netherlands: de Gruyter.

Knight, D. (2011). *Multimodality and Active Listenership*. London, UK: Continuum.

Knight, D., Evans, D., Carter, R., & Adolphs, S. (2009). HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development. *Corpora, 4*(2), 1–32.

Kress, G. (2011). Multimodal Discourse: The Modes and Media of Contemporary Communication. London, UK: Bloomsbury.

Kress, G., & van Leeuwen, T. (2001). Multimodal Discourse. London, UK: Bloomsbury.

Loehr, D., P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology, 3(1), 71–89.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. Chicago, IL: University of Chicago Press.

Norris, S. (2008). Some thoughts on personal identity construction. In V. Bhatia, K. J. Flowerdew & R. H. Jones (Eds.), Advances in Discourse Studies (pp. 132–148). London, UK: Routledge

O'Halloran, K., L. (2008). Systemic functional-multimodal discourse analysis (SF-MDA): Constructing ideational meaning using language and visual imagery. Visual Communication, 7(4), 443–475.

O'Halloran, K. L. (2011). Multimodal discourse analysis. In K. Hyland & B. Paltridge (Eds.), Continuum Companion to Discourse Analysis (pp. 120–137). London, UK: Continuum.

O'Toole, M. (1990). A systemic-functional semiotics of art. Semiotica, 82(3/4), 185–209.

Olsher, D. (2004). Talk and gesture: the embodied completion of sequential actions in spoken interaction. In R. Gardner & J. Wagner (Eds.), Second Language Conversations (pp. 221-245). London, UK: Continuum.

Schegloff, E., A. (2007). Sequence Organization in Interaction: A Primer in Conversation Analysis, Volume 1. Cambridge, UK: Cambridge University Press.

Schegloff, E., A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language, 53(2), 361–382.

Seo, M.-S., & Koshik, I. (2010). A conversation analytic study of gestures that engender repair in ESL conversational tutoring. Journal of Pragmatics, 42(8), 2219–2239.

Taleghani-Nikazm, C. (2015). On multimodality and coordinated participation in second language interaction: A conversation-analytic perspective. In D. Koike, A. & C. Blyth, S. (Eds.), Dialogue in Multilingual and Multimodal Communities (pp. 79–103). Amsterdam, Netherlands: John Benjamins.

Tsuchiya, K. (2013). Listenership Behaviours in Intercultural Encounters: A Time-Aligned Multimodal Corpus Analysis. Amsterdam, Netherlands: John Benjamins.

van Leeuwen, T. (2005). Introducing Social Semiotics. London, UK: Routledge.

van Leeuwen, T., & Jewitt, C. (Eds.). (2001). Handbook of Visual Analysis. London, UK: Sage.

Ute Römer

Corpus research for SLA

The importance of mixing methods

Abstract: This chapter argues that corpus linguistics has a lot to offer to research and practice in Second Language Acquisition (SLA), especially if different methods and data types are combined in a "methodological pluralism" sense (McEnery & Hardie 2012: 227). The chapter also suggests that progress in corpus-based SLA research will depend to some extent on successful collaborations between corpus linguists and scholars from related fields. After a brief overview of some existing uses of corpora in SLA, the chapter will present findings from two case studies that benefited from mixing methods and from the author's collaboration with researchers from neighbouring disciplines.

1 Introduction: Corpora and SLA

Over the past few decades, the growing availability of native speaker and learner corpora has enabled Second Language Acquisition (SLA) researchers to study patterns in the linguistic input of learners, as well as in their language output in a more empirical and systematic way than previously possible. Corpus linguists have contributed considerably to a better understanding of central aspects of second language (L2) learners' production and of differences between learner and native speaker English. I am here particularly thinking of the influential work of Sylviane Granger and her team at the Université Catholique de Louvain, Belgium, and of research carried out by Stefan Gries and collaborators (see e.g., Granger 2009; Granger, Gilquin, & Meunier 2013; Granger, Petch-Tyson, & Hung 2002; Gries 2008; Gries & Deshors 2014; Gries & Wulff 2005, 2009; Paquot & Granger 2012). This chapter serves to summarize some recent corpus work with clear implications for SLA, while highlighting the importance that combining research methods plays in this context.

My starting point for this contribution is the claim that the field of SLA needs corpora and corpus researchers. Empirical evidence is required to address the types of questions SLA researchers are interested in. Some of the core things that

Ute Römer, Georgia State University, uroemer@gsu.edu

https://doi.org/10.1515/9783110489071-018

SLA research aims to better understand are (i) the process of second language acquisition, (ii) developmental stages in SLA, (iii) the input learners receive and the effect it has on them, (iv) the linguistic choices made by learners and how they differ from those made by first language (L1) speakers, (v) patterns in learner production data, and (vi) potential effects of learner age, L1, other languages, motivation, etc. on acquisition. I would argue that corpus research can contribute to a better understanding of all these issues and that native speaker and learner corpora can be considered an ideal data source for SLA researchers. As Lozano and Mendikoetxea (2013: 66) point out, "[m]uch SLA research has traditionally relied on elicited experimental data while disfavoring natural language data". Such data types include grammaticality judgment tasks, fill-in-the-blanks exercises, and acceptability tasks. The growing availability of native speaker and learner corpora enables us to more systematically study patterns in the linguistic input of learners as well as in learner output. Since learner corpora "contain data from hundreds (sometimes thousands) of learners," they can arguably "lay claim to greater representativeness than previous SLA studies" (Granger 2009: 16).

Following the pioneering work by Granger and colleagues based on corpora such as ICLE (the International Corpus of Learner English; Granger et al. 2009) and LINDSEI (Louvain International Database of Spoken English Interlanguage; Gilquin, De Cock, & Granger 2010), recent research in corpus-based SLA has demonstrated that applying sophisticated quantitative methods to learner and native speaker corpora can result in insights into learner language processing and cognition that have not previously been available. In a study of the genitive alternation in Chinese and German L2 learners of English, Gries and Wulff (2013: 352) show how a multi-factorial, contextualized approach involving logistic regression analysis can "help us develop a more precise qualitative understanding of how native and learner English differ". Also using a logistic regression and data from learner and native speaker corpora, Wulff, Lester and Martinez-Garcia (2014) examine under which conditions L1 German and L1 Spanish learners of English tend to produce or omit the complementizer *that* in their writing. Their model suggests "that L2 learners' and natives' production is largely governed by the same factors" (2014: 271). In a similar vein, Gries and Deshors (2014) suggest a new statistical approach to L2 learner language analysis called 'MuPDAR' (Multifactorial Prediction and Deviation Analysis with Regressions) with the goal to better understand differences between learner and native speaker corpus data. Focusing on English native speakers' and French and Chinese learners' use of the modals *may* and *can*, the authors are able to answer the question 'What leads learners to making choices that are different from those made by native speakers?' (see also Deshors 2015).

In addition to using sophisticated approaches to analysing data from learner corpora, other recent corpus studies in SLA have combined corpus evidence and evidence collected in experimental settings. Littré (2015), for example, draws on data from a longitudinal corpus of written L1 French learner English as well as data collected in an interpretation task to better understand beginning learners' perception and use/misuse of the simple present and present progressive. Mollin (2014) combines corpus and grammaticality judgment data to test whether binomial reversibility (e.g. *day and night* vs. *night and day*) is a salient feature of English language use and also psychologically real in native speakers and learners of English. Her observations include that "the intuitions of non-native speakers do run parallel to corpus data for many binomials, if not as closely as those of native speakers do" (2014: 212). Other relevant mixed-data studies include Gries and Wulff (2005, 2009) in which the authors combine data on ditransitives vs. prepositional datives and gerundial vs. infinitival complement constructions, gathered from a corpus and in sentence completion tasks to show that advanced L1 German learners of English have verb-constructional knowledge similar to native speakers. These studies collectively show that different types of data can present converging evidence which in turn helps strengthen research hypotheses. I consider the studies summarized in the last two paragraphs methodologically particularly interesting and forward-looking, and see one of their main strengths in their skilful combination of data from different sources and of quantitative and qualitative analytic techniques.

In the remainder of this chapter, I will provide overviews of and share selected results from two case studies in corpus-based SLA that also combine quantitative and qualitative methods in the analysis of learner and/or native speaker production data of different types:

1. A study of attended and unattended *this* and the factors that influence its distribution in advanced student writing across disciplines (Wulff, Römer, & Swales 2012); and
2. A study that examines verb-argument constructions in language use and acquisition, drawing on corpus data and psycholinguistic evidence (Ellis, O'Donnell & Römer 2013; Ellis, Römer & O'Donnell 2016; Römer, O'Donnell & Ellis 2014, 2015; Römer, Roberson et al. 2014).

Both studies have benefited from mixing various analytic methods and data types. They have also benefited from my collaboration with researchers from neighbouring disciplines, including a psycholinguist, a computational linguist, a genre expert, and a cognitive linguist.

2 Corpus research for SLA: Two case studies

The following sections illustrate in what ways corpus methods can be beneficial in second language acquisition research.

2.1 Attended and unattended *this* in student writing

While not based on learner output, the first case study examines language that can serve as a model to learners and has clear implications for second language teaching practice. It combines quantitative and qualitative approaches to the distribution of attended and unattended *this* in successful student writing across disciplines and determines what learners need to know about *this* in this particular writing context. The study was carried out in collaboration with a cognitive linguist (Stefanie Wulff) and an academic discourse and genre expert (John Swales). A detailed account of the study can be found in Wulff, Römer and Swales (2012).

This first study addresses the question "What governs an academic writer's choice between attended and unattended *this*?" (Wulff, Römer, & Swales 2012). With a focus on methodology, the study also explores how in-depth qualitative investigations of *this* in context can be guided by results from quantitative and multifactorial analyses. It exemplifies how combined evidence from quantitative and qualitative methods can provide a much more comprehensive picture of a linguistic phenomenon than either method could achieve alone. The corpus that this case study is based on is a pre-final version of the Michigan Corpus of Upper-level Student Papers (MICUSP; O'Donnell & Römer 2012; Römer & O'Donnell 2011). This version of MICUSP has a size of around 2.3 million words and consists of 810 student writing samples from 16 different academic disciplines.

This is one of the most frequent words in academic writing, occupying rank 11 in a MICUSP word frequency list (see Römer & Wulff 2010). Still, the factors that determine whether it is attended by a noun as in (1) or free-standing as in (2) have not received much attention in corpus research.

(1) This finding indicates that our method is consistent.
(2) This indicates that our method is consistent.

Also, style guides and reference works conflict with actual usage when it comes to this language point. This has the potential of confusing learners who are working towards becoming more proficient L2 writers. Markel (2004: 229), for example, suggests that "[i]n almost all cases, demonstrative pronouns should be followed by nouns". As corpus research indicates, expert writers do, however, often use demonstrative *this* without a noun phrase. In a study based on the Hyland Corpus of published research articles, Swales (2005) reports that 36% of all occurrences of sentence-initial *this* in this type of expert academic writing were left unattended.

In order to provide clarification on this issue and help better understand when *this* is attended or left unattended, we carried out a systematic analysis of all sentence-initial instances of *this* in 810 texts from MICUSP (5,827 instances altogether). 57% (3,328) of these instances of sentence-initial *this* were attended, while 43% (2,499) were not attended by a noun or noun phrase. We carried out three types of analyses on the dataset: (i) a logistic regression analysis which helped determine the probability of attended versus unattended *this* on the basis of a set of predictor variables; (ii) a Distinctive Collexeme Analysis (DCA) which served to measure the level of distinctive association between verbs and (un)attended instances of *this*; and (iii) a phraseological pattern analysis of the most prominent *this* + verb clusters and their distribution across MICUSP disciplines, student levels, and texts.

According to the logistic regression, the strongest predictor for the distribution of (un)attended *this* in MICUSP is the lemma frequency of the verb, with high lemma frequency increasing the likelihood of attended *this*. More details on this strong lexical drive were provided by the DCA. This analysis helped us identify which items were responsible for the observed verb lemma frequency effect. Verbs that are distinctively associated with attended *this* include USE, EXAMINE, FOCUS, FIND, EXPLORE, and BASE. Among the verbs that are most distinctively associated with unattended *this* are BE, MEAN, LEAD, IMPLY, SEEM, and ALLOW – verbs that are used in the expression of evaluation, interpretation, or discussion. The latter group of verbs form clusters with the determiner (e.g., *this is*, *this means*) that show a high degree of fixedness. This means that these verbs rarely allow for an intervening noun. When the following verb form is *is*, sentence-initial *this* is unattended 63.5% of the time. With the verb form *means*, 98.4% of the instances of sentence-initial *this* are unattended (for this verb, the only attended *this* example in MICUSP is *This size-selectiveness means*...). The results of the DCA on verbs that typically appear in unattended contexts were corroborated by an n-gram extraction which highlighted a large number of fixed *this* + verb clusters in MICUSP.

The logistic regression and distinctive collexeme analyses were followed by a closer examination of a set of frequent *this* + verb clusters (*this is*, *this means*, *this leads*, *this implies*, *this seems*, and *this allows*). This included an analysis of the clusters' distribution across MICUSP disciplines and student levels, as well as their preferred textual positions (inspired by Hoey's 2005 work on textual colligation). This more qualitative view on the data pointed to a number of interesting distributional trends with respect to the use of the clusters in student papers from different disciplines, and their preference (or avoidance) to occur in a particular section of a paragraph or text. To give just one example, sentence-initial *this seems* is most frequent in Philosophy papers and, as shown in Figure 1, strongly prefers to occur in paragraph- and text-final and -medial positions. Occurrences of *this seems* towards the beginning of a paragraph or text are rare in MICUSP texts. Our textual distribution analysis of *this* + verb clusters nicely supports the DCA results on semantic groupings of verbs: combinations of *this* + verbs which are distinctively unattended mark upcoming interpretation or evaluation. The positional preferences of these combinations towards the end of paragraphs and texts reflect this trend. Correspondingly, combinations of *this* + nouns + verbs which are distinctively attended (according to the DCA) predominantly occur in text-initial position (Wulff, Römer, & Swales 2012: 149). MICUSP examples of this type of combination include *This study focuses on*... and *This paper explores*....

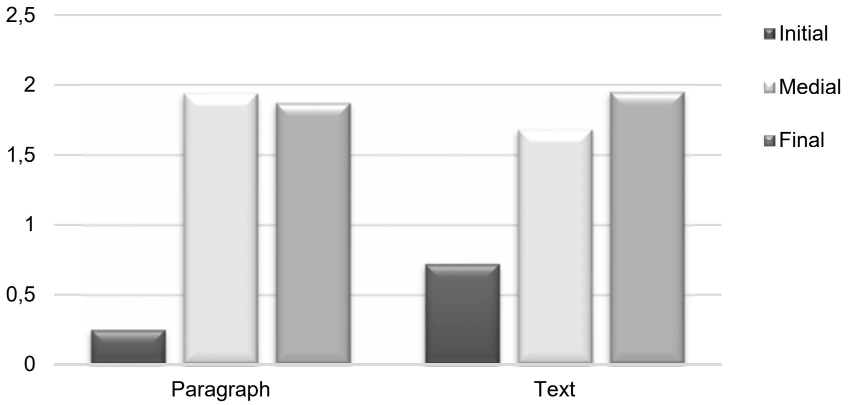


Fig. 1: Distribution of sentence-initial *this seems* across paragraphs and texts in MICUSP (figures normalized per 100,000 words).

Combining different analytic methods and different points of view on the data – a cognitive linguist's multifactorial analysis, a corpus linguist's pattern exploration, and a genre expert's qualitative analysis of *this* in context – has enabled us to identify various hitherto unexamined properties of the attended/unattended *this* alternation in proficient student academic writing. Based on our analysis, it is not just the antecedent (the focus of traditional functional studies on the topic) that determines whether *this* is attended or not, but also the following verb. Another observation we made was that sentence-initial *this* + verb clusters form fixed contiguous sequences in MICUSP which can be considered meaningful units. On the whole, the results of this case study point towards an ongoing delexicalization of *this* + verb clusters such as *this is* and *this means* into textual organization markers. Our findings are inconsistent with traditional pedagogical descriptions (e.g., Markel 2004) and their claim that unattended *this* is a mere 'vague reference' which should be avoided if at all possible. For the practice of teaching academic writing this means that advice on (un)attended *this* found in textbooks and writing manuals is clearly overgeneralized. Our study suggests that instructional materials for second language learners of English, as well as for native speakers who are novice academic writers, need to recognize that certain verb forms following *this* do not favour attending nouns. For the L2 learner or academic writing novice, it will be helpful to know that there are high frequency phrases such as *This is because* or *This means that* which need to be noted as valid exceptions to any general advice they may find in writing manuals or textbooks.

2.2 Verb-argument constructions in language acquisition and use

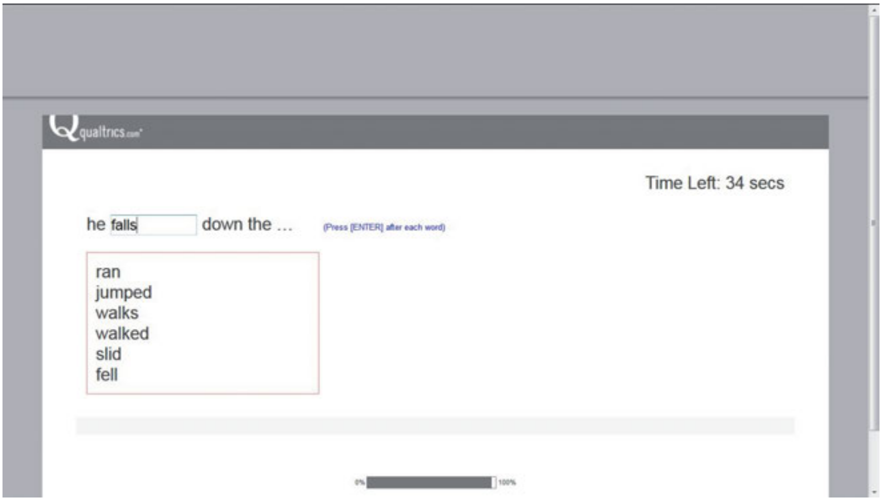
The second case study, part of a larger-scale project, examines the role that constructions play in second language acquisition. It uses data from a native speaker corpus, learner corpora, and psycholinguistic experiments to investigate what influences L2 learners' acquisition and processing of English verb-argument constructions (VACs; e.g., the 'V *about* n' construction illustrated by *he thought about her suggestion*) and what speakers know about the verbs that are most commonly associated with those constructions. The VAC project is a collaboration between a psycholinguist and second language researcher (Nick Ellis), a computational linguist (Matthew O'Donnell), and a corpus linguist (myself). Detailed accounts of different parts of the project can be found in Ellis, O'Donnell and Römer (2013, 2014a, 2014b), Ellis, Römer and O'Donnell (2016), Römer, O'Donnell and Ellis (2014, 2015), and Römer, Roberson et al. (2014).

The project takes a usage-based approach to VACs, their acquisition and processing. It addresses research questions including: 'How are verbs distributed across VACs? How are meanings created in VACs?', 'What do language users, including both native and non-native English speakers, know about VACs?' and

'What role do VACs play in first and second language acquisition?'. All collaborators brought specific analytic skills and methods to the project – methods from corpus linguistics, computational linguistics (especially natural language processing), and psycholinguistics. As a result, the project combines analyses of corpora with different types of psycholinguistic experiments.

Starting from verb patterns identified in the COBUILD Grammar Patterns (Francis, Hunston, & Manning 1996) and using tools from computational and corpus linguistics, we have mined a dependency-parsed version of the British National Corpus (BNC) for VACs. We have extracted data for around 50 VACs so far. In a team effort and working in an iterative cycle, we have defined, searched for, reviewed, and refined search patterns in order to retrieve VACs from the BNC) with highest possible precision and recall (for details on the BNC-mining procedure, see Römer, O'Donnell, & Ellis 2015). We have also extracted VAC data from corpora of spoken and written learner English, the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI), to enable comparisons of L1 and L2 speaker use of the same VACs (Römer, Roberson et al. 2014).

Psycholinguistic experiments used in this project include generative free association tasks and verbal fluency tasks. In generative free association tasks, we asked native English speakers and advanced second language learners of different L1 backgrounds (German, Czech, Spanish) to fill 40 bare VAC frames (e.g., 'she ____ about the…' or 'it ____ off the…') with the first verb that came to mind. In verbal fluency tasks, L1 and L2 speakers of English responded to the same 40 VAC frames but were asked to produce as many verbs as they could think of in one minute. The verbal fluency prompt for one of the VACs, 'V *down* n', is displayed in Figure 2. The data collected in these experiments allow for comparisons of (i) BNC usage data vs. native speaker responses, (ii) BNC usage data vs. learner responses, and (iii) native speaker responses vs. learner responses. The results of such comparisons allow us to determine in what ways usage influences native speaker and learner VAC production, and whether/how learner VAC knowledge differs from native speaker VAC knowledge.



Time Left: 34 secs

he falls down the ... (Press [ENTER] after each word)

ran
jumped
walks
walked
slid
fell

Fig. 2: Verbal fluency task prompt for the 'V *down* n' VAC.

Our large-scale BNC-based corpus analyses allowed us to demonstrate the validity of VACs in language usage. We found that VACs are Zipfian in their type-token distributions, such that a small number of verb types account for the majority of all VAC tokens while a large number of verb types only occur in a specific VAC once or twice. For instance, in the 'V *about* n' VAC, the verbs THINK and TALK are the dominant verbs, each occurring more than 3,000 times in this VAC in the BNC, whereas verbs such as QUARREL and THEORIZE, while contributing to the meaning of the construction, only have very few occurrences. The corpus analyses also indicated that VACs are selective in their verb form occupancy and coherent in their semantics (for details, see Ellis, O'Donnell, & Römer 2013).

The corpus findings allowed us to make predictions regarding language users' knowledge of verbs in constructions. We tested these predictions in psycholinguistic experiments. Through those experiments, we demonstrated the validity of VACs in native speakers' and second language learners' minds. We found that both L1 and advanced L2 speakers of English have strong constructional knowledge and that their VAC processing showed effects of usage frequency, contingency, and prototypicality (for details, see Ellis, O'Donnell, & Römer 2014a, 2014b). Comparisons of experiment data collected from learners of different first languages showed that there are also systematic differences across learner groups (L1 German vs. L1 Czech vs. L1 Spanish) in the associations of verbs and constructions. These differences can be explained on the basis of crosslinguistic transfer effects as well as effects of language typology that impact verb semantics

(Talmy 1985). Overall, our findings suggest that learners whose first language is, like English, satellite-framed (in our study German and Czech) produce more verbs that are similar to those produced by native speakers than learners whose first language is verb-framed (here Spanish; for details, see Römer, O'Donnell, & Ellis 2014).

Work we have done based on an additional, richer set of data from verbal fluency tasks and based on data retrieved from corpora of L2 learner speech and writing provides further evidence on learner VAC knowledge and how it compares to that of native English speakers. We asked three groups of participants (99 English native speakers, 96 advanced L1 Spanish, and 94 advanced L1 German learners of English) to complete VAC frames such as 'she _____ over the...' with verbs that may fill the blank. All participants produced as many verbs as they could think of in one minute and then moved on to the next VAC prompt. The VACs we selected for this set of experiments were: 'V *about* n', 'V *across* n', 'V *after* n', 'V *against* n', 'V *among* n', 'V *around* n', 'V *as* n', 'V *between* n', 'V *for* n', 'V *in* n', 'V *into* n', 'V *like* n', 'V *of* n', 'V *off* n', 'V *over* n', 'V *through* n', 'V *towards* n', 'V *under* n', and 'V *with* n'. For the same 19 VACs, we extracted data from the German and Spanish sub-sections of ICLE and LINDSEI. Frequency-sorted versions of the lemmatized verb lists for each VAC from the experiments were compared against each other and against the verb lists that resulted from the ICLE (German/Spanish) and LINDSEI (German/Spanish) analyses.

Our comparison of L2 learner and native speaker production data indicated that, while there is some overlap between learners' and native speakers' verb-VAC associations (especially for German but less so for Spanish participants), L2 learners tend to rely more on general, high-frequency verbs (including BE and DO) and produce fewer specific, less frequent verbs (such as SLIP and CRAWL). The learner corpora analyses confirmed this observation. In addition to providing us access to the verbs that are most entrenched in each VAC in the learners' minds (and which were shared with the experimental data), the datasets from ICLE and LINDSEI also highlighted other verbs that did not occur repeatedly in learners' survey responses. These verbs often depended on the types of texts included in the learner corpora and on the tasks learners were asked to perform (e.g. ARGUE and WORRY in 'V *about* n' in ICLE; see Table 1 for lists of the most frequent verbs learners used in this VAC in the two learner corpora). The experimental data on the other hand provided us with additional sets of verbs which are semantically related to the most frequent verb(s) in a VAC (e.g. SPEAK and ASK for 'V *about* n'). Both types of data (corpus and experimental) allowed us to identify the lead verb(s) in each construction (e.g. TALK for 'V *about* n').

Tab. 1: Top ten verb choices for ‘V *about* n’ across ICLE and LINDSEI datasets (Ger = L1 German learners; Spa = L1 Spanish learners).

Rank	ICLE_Ger		LINDSEI_Ger		ICLE_Spa		LINDSEI_Spa	
1	THINK	67	TALK	40	TALK	49	TALK	23
2	TALK	36	THINK	38	THINK	37	THINK	20
3	CARE	17	BE	26	CARE	8	BE	14
4	FORGET	13	COMPLAIN	8	BRING	7	SPEAK	9
5	COMPLAIN	12	KNOW	6	SPEAK	7	COMPLAIN	5
6	KNOW	8	WORRY	6	WORRY	7	ARGUE	3
7	LEARN	7	LIKE	3	FORGET	6	SAY	3
8	BRING	6	SAY	3	KNOW	6	WORRY	3
9	HEAR	6	CARE	2	HEAR	5	CHOOSE	2
10	BE	5	LAUGH	2	BE	4	HEAR	2

One issue that became apparent in this comparative study that combined different methods and data types to uncover what L2 learners know about verbs in common English VACs was that the learner corpus subsets yielded robust token (and hence type) numbers for only a few of the 19 VACs and fairly small numbers for all others (for details, see Römer, Roberson et al. 2014). For those VACs, larger token numbers would be essential if we wanted to identify semantic patterns or even lead verbs. This calls for larger learner corpora that may complement resources such as ICLE and LINDSEI. Large amounts of texts produced by L2 learners of a range of L1 backgrounds and proficiency levels have recently been made available by the Education First (EF) research unit at the University of Cambridge, UK in the EF-Cambridge Open Language Database (EFCAMDAT; see Geertzen, Alexopoulou, & Korhonen 2013). We have retrieved sets of texts produced by German and Spanish learners at CEFR levels A1 through C2 from EFCAMDAT and are now extracting VAC usage data from those text collections. The EFCAMDAT subsets we retrieved – over 28,000 texts and 2.8 million words from L1 German learners, and over 40,000 texts and 3.2 million words from L1 Spanish learners – should provide us with more robust token numbers for most VACs and will allow us to study learners’ language development.

The insights into speaker knowledge and use of VACs summarized here, and discussed in more detail in the various studies referenced throughout this section, have been gained through combining methods and data sources from corpus, computational, and psycholinguistics. These insights would not have been possible to achieve with one method or data source alone.

3 Conclusion

The aims of this chapter have been to make a case for more corpus-based research activity that has implications for Second Language Acquisition theory and practice, and to argue that progress in corpus-based SLA will depend to some extent on successful combinations of research methods and data types. This was done on the basis of a brief overview of recent relevant mixed-methods research and a review of two case studies: (1) a study of attended and unattended *this* in proficient student writing, and (2) a study of the use and acquisition of English verb-argument constructions. These studies address SLA questions with the help of native speaker and learner corpora as well as data from psycholinguistic experiments. The methods that these studies benefited from include logistic regression analysis, distinctive collexeme analysis, pattern examination, learner corpora and native speaker corpus mining for constructions, generative free association tasks, and lexical production tasks.

Aside from mixing methods, both case studies reported on here have also benefited from collaboration with colleagues from neighbouring disciplines: a psycholinguist, a computational linguist, a genre expert, and a cognitive linguist. In the first case study, the bringing together of quantitative and qualitative methods that were brought to the table by my collaborators and myself resulted in converging findings and aspects of the use of *this* in academic writing that had not been reported in previous publications on the topic. The findings had implications for second language teaching practice. In the second case study, collaboration inspired a mixed-methods approach to investigating what native speakers and learners of English know about verbs in constructions and how usage influences construction acquisition. The combination of corpus and experimental methods has produced novel results on the distribution of verbs across VACs and on the factors that influence speaker VAC knowledge.

I hope that this overview has shown that mixed-methods, collaborative work has benefits for SLA research and can lead to insights that would be hard to achieve without it. I agree with McEnery and Hardie who have argued that, in corpus linguistics, “the way ahead is methodological pluralism” (2012: 227), and with others who have called for a higher level of methodological inclusivity in corpus work (Arppe et al. 2010; Ellis & Simpson-Vlach 2009; Gilquin 2007; Gilquin & Gries 2009; Gries 2013; Littré 2015; Wulff 2009; Wulff et al. 2009). I believe that this methodological inclusivity is particularly important in the context of corpus-based SLA work which depends on a better understanding of learners’ language input, their output, and on insights into learners’ mental representations of language.

References

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora, 5*(1), 1–27.

Deshors, S. (2015). A multifactorial approach to linguistic structure in L2 spoken and written registers. *Corpus Linguistics and Linguistic Theory*, *11*(1), 19–50.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning, 63*(Supp. 1), 25–51.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014a). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics, 25*(1), 55–98.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014b). Second language processing of verb–argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism, 4*(4), 405–431.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar* (Language Learning Monograph Series). Malden, MA: Wiley.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory, 5*(1), 61–78.

Francis, G., Hunston, S., & Manning, E. (1996). *Grammar Patterns 1: Verbs*. London, UK: Harper-Collins.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). *Proceedings of the 31st Second Language Research Forum (SLRF)*. Carnegie Mellon University, Pittsburgh, PA: Cascadilla Press.

Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik, 55*(3), 273–291.

Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). *LINDSEI: Louvain International Database of Spoken English Interlanguage*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1), 1–26.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 13–32). Amsterdam, Netherlands: John Benjamins.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *ICLE: International Corpus of Learner English*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2013). *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Petch-Tyson, S., & Hung, J. (Eds.). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam, Netherlands: John Benjamins.

Gries, S. Th. (2008). Corpus-based methods in analyses of Second Language Acquisition data. In P. Robinson & N. C. Ellis (Eds.). *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 406–431). London, UK: Routledge.

Gries, S. Th. (2013). Data in Construction Grammar. In G. Trousdale & T. Hoffmann (Eds.). *The Oxford Handbook of Construction Grammar* (pp. 93–108). Oxford, UK: Oxford University Press.

Gries, S. Th., & Deshors, S. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora, 9*(1), 109–136.

Gries, S. Th., & Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics, 3*, 182–200.

Gries, S. Th., & Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics, 9*, 163–186.

Gries, S. Th., & Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners. *International Journal of Corpus Linguistics, 18*(3), 327–356.

Hoey, M. P. (2005). *Lexical Priming: A New Theory of Words and Language*. London, UK: Routledge.

Littré, D. (2015). Combining experimental data and corpus data: Intermediate French-speaking learners and the English present. *Corpus Linguistics and Linguistic Theory, 11*(1), 89–126.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition. The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.). *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65–100). Amsterdam: John Benjamins.

Markel, M. (2004). *Technical Communication* (7th ed.). Boston, MA: Bedford/St. Martins.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics. Method, Theory and Practice*. Cambridge, UK: Cambridge University Press.

Mollin, S. (2014). *The (Ir)reversibility of English Binomials: Corpus, Constraints, Developments*. Amsterdam, Netherlands: John Benjamins.

O'Donnell, M. B., & Römer, U. (2012). From student hard drive to web corpus (part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 7*(1), 1–18.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics 32*, 130–149.

Römer, U., &, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 6*(2), 159–177.

Römer, U., O'Donnell, M. B, & Ellis, N. C. (2014). Second language learner knowledge of verb-argument constructions: Effects of language transfer and typology. *The Modern Language Journal, 98*(4), 952–975.

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge. In N. Groom, M. Charles, & S. John (Eds.). *Corpora, Grammar and Discourse: In Honour of Susan Hunston (*pp. 43–71). Amsterdam, Netherlands: John Benjamins.

Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal, 38,* 115–135.

Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research, 2*(2), 99–127.

Swales, J. M. (2005). Attended and unattended “this” in academic writing: A long and unfinished story. *ESP Malaysia, 11*, 1–15.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical form. In T. Shopen (Ed.). *Language Typology and Syntactic Description: Grammatical Categories and the Lexicon* (pp. 57–149). Cambridge, UK: Cambridge University Press.

Wulff, S. (2009). *Rethinking Idiomaticity. A Usage-based Approach*. London, UK: Continuum.

Wulff, S., Ellis, N. C., Römer, U., Bardovi-Harlig, K., & LeBlanc, C. (2009). The acquisition of tense-aspect: Converging evidence from corpora, cognition, and learner constructions. *The Modern Language Journal, 93*(3), 354–369.

Wulff, S., Lester, N., & Martinez-Garcia, M. T. (2014). *That*-variation in German and Spanish L2 English. *Language and Cognition, 6*(2), 271–299.

Wulff, S., Römer, U., & Swales, J. M. (2012). Attended/unattended *this* in academic student writing: Quantitative and qualitative perspectives. *Corpus Linguistics and Linguistic Theory, 8*(1), 129–157.

List of contributors

Helen Baker is a Research Fellow at the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University. She is a historian whose research focuses on the benefits of using large corpora in the study of the past.

Beatrix Busse is Professor of English Linguistics at Heidelberg University. Her research interests include the history of English, English historical linguistic and Shakespeare studies, stylistics and corpus linguistics. She is the Co-editor of the series *Discourse Patterns* and Reviews Editor of the *International Journal of Corpus Linguistics*.

Dagmar Deuber holds the Chair of Variation Linguistics at the University of Münster, Germany. Her research is devoted to varieties of English world-wide, with a special focus on the Caribbean, and uses a variety of approaches, in particular corpus linguistics and sociolinguistics.

Hildegunn Dirdal is Associate Professor of English Language at the University of Oslo. Her main research interests lie in the fields of translation studies, contrastive linguistics and second language acquisition.

Gregory Garretson is a Senior Lecturer in English Linguistics in the Department of English at Uppsala University, Sweden. His interests include corpus linguistic theory and methods, corpus compilation, and the development of corpus tools, especially for the study of semantics and discourse.

Anne-Line Graedler is Professor of English/English Didactics at Inland Norway University of Applied Sciences. Her main research interests are different aspects of learner English at various levels, and language contact and English influence on Norwegian, primarily based on corpus linguistics.

Ian Gregory is Professor of Digital Humanities at Lancaster University where he founded their Digital Humanities Hub. His primary interest is in how geospatial technologies can be used in humanities research, particularly with textual sources.

Eva Canan Hänsel works at the chair as a Research Assistant and is a student at Münster University's Graduate School of Empirical and Applied Linguistics. Her PhD research is situated in the area of English in the Caribbean.

Daniel Hartmann is a GIS Professional working for the RPS group. His current interests include reducing flood risk in urban areas.

Hilde Hasselgård is a Professor of English Language and Linguistics at the University of Oslo. Her present research focuses mainly on corpus-based contrastive analysis and learner corpus research, particularly at the interfaces between lexis, grammar and discourse.

Mark Kaunisto is a Senior Lecturer at the Faculty of Communication Studies in the Degree Programme of English Language, Literature and Translation at the University of Tampere.

Rolf Kreyer studied English Language and Literature, Mathematics and Education Studies at the University of Bonn. Currently he holds the Chair of English Linguistics at the University of Marburg. His research interests include learner corpus linguistics, syntax, text linguistics and cognitive linguistics.

Tove Larsson's research interests include corpus linguistics, L2 writing and syntax. She has a PhD in English linguistics from Uppsala University in Sweden.

Magnus Levin is Associate Professor of English Linguistics at Linnaeus University, Sweden. His research interests include change and variation in different varieties of English. More recent research areas concern English used in Sweden and contrastive linguistics.

Michaela Mahlberg is Professor of Corpus Linguistics at the University of Birmingham, UK, where she is also the Director of the Centre for Corpus Research. Michaela is the Editor of the *International Journal of Corpus Linguistics* (John Benjamins) and together with Wolfgang Teubert she edits the book series *Corpus and Discourse* (Bloomsbury).

Tony McEnery is a Distinguished Professor of English Language and Linguistics at Lancaster University. He has worked with scholars from a broad range of subjects and an array of impact partners including British Telecom, the Home Office and IBM.

Alessandra Molino is Lecturer of English Linguistics and Translation at the University of Torino, Italy. Her research interests focus on English-medium instruction, academic discourse and business discourse on sustainability, three areas which she explores combining the tools of corpus linguistics and discourse analysis.

Susan Nacey is Professor of English as a Second/Foreign Language at Inland Norway University of Applied Sciences, where she is Vice Dean for Research. She primarily researches metaphor in discourse and is the author of *Metaphors in Learner English* (John Benjamins, 2013).

Ute Römer is an Associate Professor of Applied Linguistics at Georgia State University. Her research interests include corpus linguistics, phraseology, and second language acquisition. She is General Editor of the *Studies in Corpus Linguistics* book series (John Benjamins).

Sylvi Rørvik is Associate Professor of English Language at Inland Norway University of Applied Sciences. Her primary research interests are in the field of learner corpus research, focusing particularly on argumentative writing, academic writing, and factors influencing language learners' proficiency.

Juhani Rudanko is Professor Emeritus of English Philology at the University of Tampere. He is the author of some 70 peer reviewed books and articles, including studies on variation and change in the system of English predicate complementation.

Wolfgang Teubert, since 2013 retired from the University of Birmingham, has always been fascinated by the notion of meaning. He aims to establish hermeneutics as the theoretical framework of discourse analysis, with corpus linguistics as its methodology. Language, for him, is not a mechanical system but exists in form of discourse.

Keiko Tsuchiya is Associate Professor at the International College of Arts and Sciences, Yokohama City University, Japan. Her research interests include multimodal corpus analysis, healthcare interaction, and English as a Lingua Franca in institutional and academic settings.

Viola Wiegand is a Research Fellow at the University of Birmingham. Her main research interest is the study of textual patterns in relation to meaning with frameworks from corpus linguistics, stylistics and discourse analysis. She is Assistant Editor of the *International Journal of Corpus Linguistics*.

Leonie Wiemeyer is a Programme Manager at Bremen Early Career Researcher Development and coordinates the writing workshop of the Faculty of Linguistics and Literary Studies at the University of Bremen. In her PhD project, she investigates the processes and products of source-based academic writing in L2 English.

Index

https://doi.org/10.1515/9783110489071-019