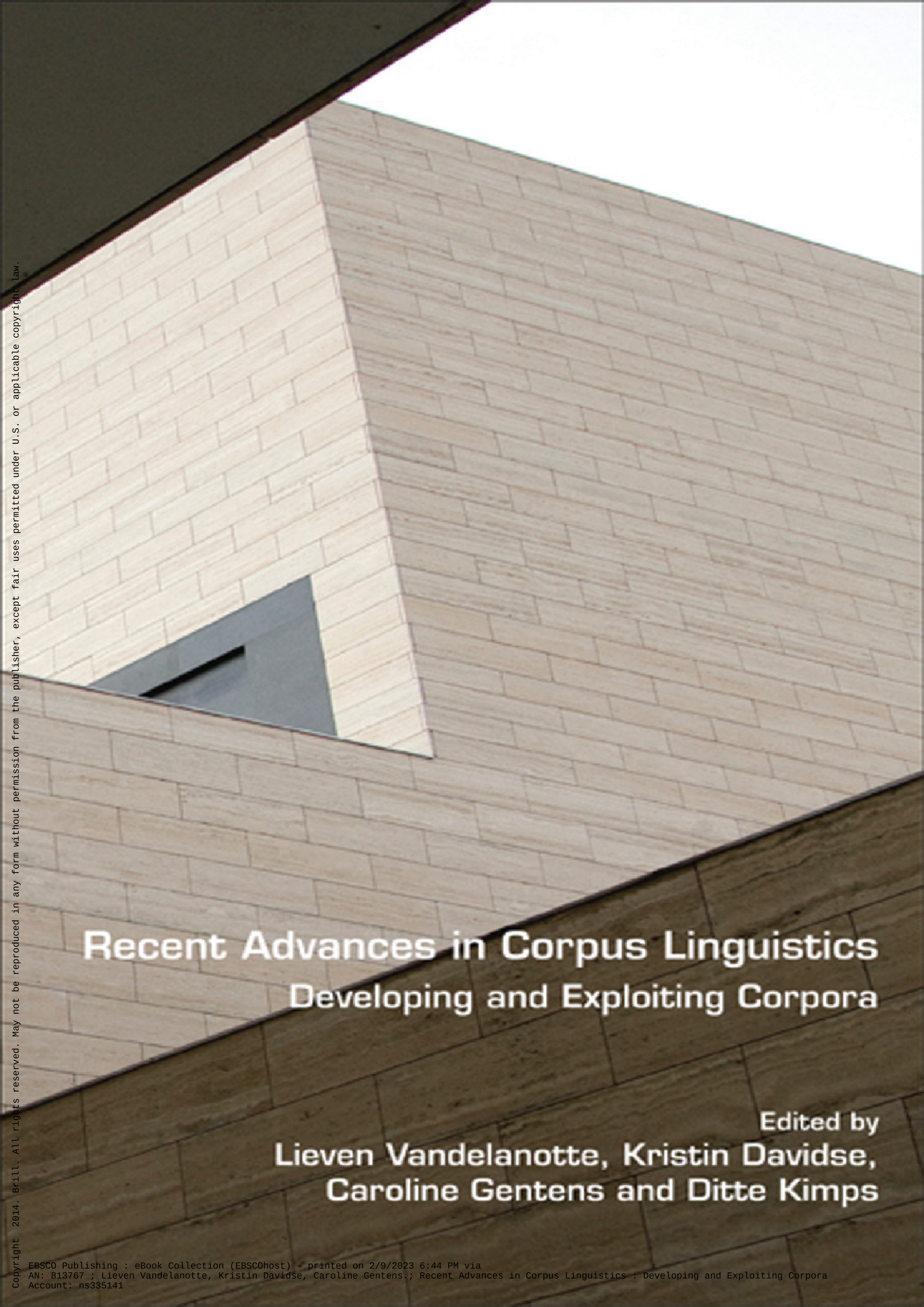


Recent Advances in Corpus Linguistics

Developing and Exploiting Corpora

Edited by
Lieven Vandelanotte, Kristin Davidse,
Caroline Gentens and Ditte Kimps

Recent Advances in Corpus Linguistics

LANGUAGE AND COMPUTERS: STUDIES IN PRACTICAL LINGUISTICS

No 78

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

Recent Advances in Corpus Linguistics

Developing and Exploiting Corpora

Edited by

Lieven Vandelanotte
Kristin Davidse
Caroline Gentens
Ditte Kimps

Rodopi

Amsterdam - New York, NY 2014

Cover image: photo of M – Museum Leuven taken during the 33rd ICAME conference by Sebastian Hoffmann.

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-3871-4
E-Book ISBN: 978-94-012-1113-0

Printed in The Netherlands

To the memory of
Monique J. van der Haagen
(1956–2012)

Contents

Part 1. Corpus development and corpus interrogation

Part 2. Specialist corpora

Part 3. Second language acquisition

Acknowledgements

This volume is a selection of strictly refereed and extensively revised papers from the ICAME 33 International Conference "Corpora at the Centre and Crossroads of English Linguistics", organised in Leuven from 30 May to 3 June 2012. The conference was convened by Kristin Davidse (University of Leuven) and Lieven Vandelanotte (University of Namur). They were very pleased that Caroline Gentens and Ditte Kimps joined the editorial team.

While naturally we accept the final responsibility for the choices made, we were very lucky in being guided by the conscientious advice of the following external referees: Karin Aijmer, Bengt Altenberg, John Bateman, Lennart Bierkandt, Rubén Chacón-Beltrán, Kees de Bot, Sylvie De Cock, Pieter de Haan, Guy De Pauw, Lynne Flowerdew, Diana Frantzen, Jean-Michel Fournier, Ulrike Gut, Raymond Hickey, Elise Kärkkäinen, Przemek Kaszubski, Merja Kytö, Kerstin Lunkenheimer, Michael MacMahon, Ilka Mindt, Nadja Nesselhauf, Monika Reif, Tom Ruette, Anne-Marie Simon-Vandenbergen, Rob Spence, Benedikt Szmrecsanyi, Johan Strobbe, Irma Taavitsainen, Paul Tench, Vincent Vandeghinste and Kris Van den Branden. We much appreciated their careful comments and suggestions for revision, which benefited both the papers selected for this volume and those eventually recommended to other publication outlets.

Without the contributors, there obviously would not shave been a volume. We thank them for developing the studies they presented at ICAME 33 with an eye to the general coherence of this volume and for the good spirit in which the whole enterprise was conducted.

Rodopi helped along the project with their usual friendly efficiency. Eric Van Broekhuizen was always ready to offer help and useful advice and we are very grateful to the series editors Christian Mair, Charles Meyer and Nelleke Oostdijk for having accepted the volume for publication in the series *Language and Computers: Studies in Practical Linguistics*.

For generous financial support of the conference, thanks are due to the Research Foundation – Flanders (FWO), the *Fonds de la Recherche Scientifique* (FNRS) and the Leuven International Doctoral School for the Humanities and Social Sciences. We would also like to acknowledge the support of the GOA-project 12/007, *The multiple functional load of grammatical signs*, awarded by the Leuven Research Council and coordinated by Kristin Davidse, Bert Cornillie, An Van linden and Jean-Christophe Verstraete.

Introduction

Kristin Davidse, Caroline Gentens*, Ditte Kimps* and Lieven Vandelanotte***

*KU Leuven – University of Leuven
**University of Namur

This book is a collection of contributions that represent recent advances in corpus linguistics – both in the development of specialist corpora and in ways of exploiting them for specific purposes. The volume is a selection of studies that were presented at the ICAME 33 International Conference “Corpora at the Centre and Crossroads of English Linguistics” (Leuven, 30 May - 3 June 2012). As reflected in the title, the conference had a dual focus. On the one hand, it focused on the importance of optimal compilation and analysis of corpus data for the ‘core business’ of English linguistics. It is now generally accepted that the study of English requires optimally compiled data collections and the development of methodologies to interrogate the data. By the same token, the very questions linguists ask have been moulded and changed by the availability of large corpora. On the other hand, the conference focused on the ‘intersections’ between English corpus linguistics and other domains of linguistics. The conference wanted to offer a forum for ‘crossroads’ to which corpus data contribute extra value and which, in their turn, may open new perspectives on data compilation and analysis. The intersections of longest standing are (i) educational linguistics, in which learner corpora have been innovative and are further evolving (e.g. from written to spoken learner language); (ii) language acquisition, which puts very specific demands on the collection of language data; (iii) contrastive linguistics, which requires parallel corpora.

ICAME 33 more than realized these aims. The conference was a successful and stimulating event at which all the above themes were explored and existing boundaries pushed forward. It brought together researchers from all generations of the extended ICAME family with its variegated corpus-oriented interests as well as researchers who were new to these meetings. The papers in this volume have been grouped into the thematic clusters of corpus development and automated corpus interrogation, studies based on specialist corpora, and second language acquisition studies. Within each thematic cluster new challenges are tackled and novel ways of getting mileage out of specialist corpora are reported on.

The first part, ‘Corpus development and corpus interrogation’, opens with two papers describing the challenges involved in the development of two highly specialized corpora that will fill important gaps in historical databases. **Auer, Laitinen, Gordon and Fairman** report on the making of an electronic corpus of letters written between c. 1750 and 1835 by English men and women from the lowest social strata, *Letters of Artisans and the Labouring Poor* (LALP). They give a detailed description of the collection, transcription and coding procedures they applied. Their main goal is to make the corpus accessible and searchable not only for linguists, but for sociologists and historians as well. The corpus will be available in different formats, focusing not only on textual transcription, but also on visual representations and meta-textual information. **Beal and Sen** set out the main lines of a project to develop a corpus of eighteenth-century English phonology, based on eighteenth-century pronouncing dictionaries and other texts dealing with pronunciation. The database will be made up of Unicode transcriptions of lexical word sets with links to descriptive and prescriptive comments made in the primary sources. It will also include meta-textual information useful for research investigating phonological variation relating to time, space and social class. The authors present a pilot study demonstrating how such a database can be used to answer questions concerning the chronological, social, geographical and phonological distribution of variation between /hw/ ~ /w/ ~ /h/ in words such as WHICH, WHO and NOWHERE.

The remaining papers in the first part attack from various angles the question of how relevant patterns can be extracted automatically from computerized datasets. **Garretson and Kaatari** present the "shared evaluation" approach as an alternative to typical concordancer-based approaches, in which thanks to an in-built scoring system a greater share of the sorting and classifying of corpus data is taken over by the computer, but in which a human coder ensures accuracy in a subsequent step. As an example of one specific program which implements this approach, the authors explain the basic architecture of their program SVEP, and a case study on adjective complementation using SVEP illustrates the main merits and potential drawbacks of both the specific program, which is freely available to researchers, and of the shared evaluation approach more generally. **Schilk** uses the *International Corpus of English*, whose corpus design includes a large number of spoken texts and a large variety of different text-types and genres, as the basis for the description of text-type differences in World English. He creates currency annotated part-of-speech tag profiles of the different subcorpora and of their text-types and genres. He then uses these profiles to identify homogeneous text-type groups, finding, amongst others, that the text types 'written' and 'spoken' are best defined as conceptually, rather than medially, written or spoken. He detects corpus design artefacts and ends on a plea for more consistent corpus design. **Zumstein** demonstrates that word-stress variation can be investigated, relying on a sub-corpus of entries with stress-variants that were extracted from a computer-searchable version of Wells's *Longman Pronunciation Dictionary* (2009). The specific question that he seeks to answer is if word-stress variants in lexicophonetic corpora are exceptional cases or regular forms. Following Guierre's (1966a, 1966b, 1979) theory of word stress, he proposes that word-stress variation in English is the result of conflicting rules and forms the locus of ongoing changes. The directions of the changes are determined with the help of diachronic data extracted from pronouncing dictionaries of the eighteenth and nineteenth centuries. Based on a set of case studies, the article shows that Guierre's Normal Stress Rule is often a regularising force when competing with minor stress rules.

The second part consists of descriptive studies that draw on ‘Specialist corpora’. The first two articles investigate specific grammatical patterns in World Englishes as attested in the *International Corpus of English*. **Collins, Yao and Borlongan** examine short-term diachronic changes in the use of relativizers in Philippine English. They find that *that*-relatives are on the rise at the expense of *wh*-relatives in Philippine English, which is in line with Leech et al.’s (2009) findings for British and American English. Further examination of regional, stylistic and genre variation shows that the development reflects further alignment of Philippine English with its ‘colonial parent’, American English, which argues for Philippine English’s not yet having reached the stage of ‘endonormative stabilization’ (Schneider 2007). **Schilk and Hammel** reveal internal variation within the general ‘overuse’ of the progressive in South Asian and Southeast Asian varieties of English: in variety-based clusters, Sri Lankan English is grouped together with the Southeast Asian rather than the South Asian cluster in terms of progressive aspect marking, and Singapore English does not form part of any cluster. This variation in the use of the progressive is shown to be partly due to socio-historic factors (with Singapore English being the endonormatively stabilized variety) or to medium and text-type based differences. Varieties may also differ in their range of innovative uses of the progressive with particular verbs, for instance concerning tense marking, so-called ‘stacked progressives’, different collocational profiles or overextension to stative contexts. The use of stative verbs in the progressive, however, was found to be less influential than has often been claimed, as it turned out to be a factor mainly in the spoken data. **Renouf** develops a new perspective on neology on the basis of qualitative and quantitative profiles drawn from the 1.3 billion-word diachronic corpus of UK *Guardian* news texts, published between 1984 and 2012. She argues that one or a number of specific coinages associated with a major topical event will spread to a broader set of associated uses, thereby causing an incipient, possibly ephemeral kind of ‘register’ to emerge in a discourse community. Corpus analysis of *Guardian* news texts from the second half of 2011 tracks across time how initial coinages such as *big society* or *squeezed middle* lead to reactive neologising (as when *small society* or *alarm clock Britain* are coined as other politicians’ responses to *big society* and *squeezed middle*). This is accompanied by existing (often specialist) terms increasing in frequency (*empowering, stakeholder*), taking on new meanings (*deliver public services*) or grammatical patterns (*to impact the UK*), and co-occurring and converging in meaning (*outreach and stakeholder engagement*). **Egan and Rawoens** examine the hypothesis that translation equivalents may be employed to cast light on the semantic network of a lexeme in its original language, which is effectively a variant of the semantic mirrors method in contrastive corpus linguistics. They focus on the preposition *amid(st)* and *among(st)*, commonly taken to overlap in meaning, which they investigate in the English language original texts that are common to both the *English-Norwegian Parallel Corpus* and the *English-Swedish Parallel Corpus*. They provide new insights into the semantics of *amid(st)* and *among(st)*, establishing the senses in their semantic networks that are most frequent, most central, and that have most connections to other meanings in the network. **Kunz and Lapshinova-Koltunski** tackle the question of how the similarities and differences between the use of cohesive (i.e. non-structural) connectives in English and German can be captured. They first discuss a general classification of types of connectives and their system-related properties in each language. They then proceed to a corpus linguistic investigation of how connectives occur in original English and German texts as well as in translations in the *German English Corpus of Translations and Originals* (GECCo). Their first findings reveal differences in the textual realizations in terms of frequencies and functions.

The third part brings together five studies in 'Second language acquisition'. **Deroey** offers one of the first studies of markers of relevance and lesser relevance in lectures, taken from the corpus of *British Academic Spoken English*. She shows that they combine discourse organisation with evaluation: they help students discern the relative importance of points, thus aiding comprehension, note-taking and retention. For markers of relevance, as in *remember that…, the bottom line is…* or *it's important to point out that…*, a categorisation is presented in terms of lexicogrammatical patterns in the vein of Pattern Grammar (Hunston and Francis 2000). Markers of lesser relevance are less easily defined in such terms, as they depend quite strongly on discourse context for interpretation, which is why a pragmatic categorisation is developed using categories such as message status (*as an aside*), topic treatment (*we won't go into the detail*) and lecturer knowledge (*I can't remember the exact details*). In addition to its value for EAP teaching and learning, a number of unexpected or less predictable findings demonstrate the usefulness of the corpus approach, while Deroey also considers possible limitations of using transcripts only, without recourse to recordings or interviews with discourse participants. **Roca-Varela** tackles deceptive cognates in the written and spoken language of Spanish learners of English, as documented in the *International Corpus of Learner English* (ICLE) and the *Louvain International Database of Spoken English Interlanguage* (LINDSEI). Her quantified corpus study reveals that certain English false friends remain problematic with Spanish learners (e.g. *casualties, facilities, argument, career, rare*) because of their formal resemblance to Spanish words with a different meaning, even though the wrong uses may cause misunderstandings. She also finds that, interestingly, learners make more errors with false friends in their written than in their spoken production; *realize* and *actual*, for instance, are used with more idiomatic and collocational control in speech than in writing. The contribution by **Gaillat**, **Sébillot and Ballier** is part of an ambitious project to develop a fine-grained automatic annotation process that will be applicable to any corpus. The present article relies on NLP tools to classify instantiations of native and non-native uses of the demonstratives *this* and *that*. The native uses were extracted from the Penn Treebank-tagged *Wall Street Journal corpus*, and the non-native ones from the English learner corpus *Charliphonia*, the University of Paris-Diderot's subset of the *Longdale* corpus. Based on the token context and their PoS tags, the tools identify which class of features are decisive for expected (preferred) or unexpected (second-choice/erroneous) uses.

The volume ends with two studies initiated by van der Haagen and de Haan. The first by **van der Haagen, de Haan and de Vries** is a pilot study into the feasibility of charting student speakers' spoken proficiency, based on successive recordings of unprepared speech by students of English at Radboud University Nijmegen. It looks at the often assumed but under-researched link between students' actual proficiency and a specific level of the CEFR (Common European Framework of Reference for Languages), and the question of how students' progress in language proficiency can be correlated with a higher CEFR level. As a first attempt at finding out whether or not it is possible to measure students' progress in spoken English over the first two years of their degree course objectively, they undertake a small pilot study of a corpus of spoken English which they collected, involving 31 participants from a single cohort who were recorded in their first week at university in the first and second year. Somewhat surprisingly, they found that lexical density decreased, which might however be an indirect measure of syntactic advancement related to what hearers can process. The second by **de Haan and van der Haagen** is a longitudinal study of the syntactic development in EFL writing of very advanced Dutch students of English at the same university. They explore the possibilities of a longitudinal study which tracks how the writing of advanced learners of English develops over time. They collected data from a single cohort of students of English at Radboud University Nijmegen, who started their studies in 2011. Based on this – modest – amount of longitudinal data, they offer a quantitative and a qualitative analysis of how non-native writing develops over time, whether it develops in the direction of native writing, and whether individual students display individual developmental patterns. The answer to all three questions appears to be affirmative.

A few months after the ICAME 33 conference, Monique van der Haagen, who co-authored two of the papers collected here, fell seriously ill, and unfortunately her illness progressed unstoppably. True to her spirit, Monique celebrated life until she had to let go of it, on 17 October 2012. On behalf of the broader ICAME community, the editors would like to dedicate this volume to Monique's memory, as a token of our deep respect for her scholarship as much as for her affability, good cheer and wit, for which she is warmly remembered.

References

Guierre, L. (1966a), 'Éléments pour une étude linguistique de l'accentuation en anglais', *Les Langues Modernes*, 1: 161-170.

Guierre, L. (1966b), 'Traitement automatique des langues: un codage des mots anglais', *Études de Linguistique Appliquée*, 4: 48-63.

Guierre, L. (1979), *L'Accentuation en Anglais Contemporain, Éléments pour une Synthèse*, Paris: Département de Recherches Linguistiques, Institut d'anglais Charles 5. Université de Paris 7 - Denis Diderot.

Hunston, S. and G. Francis (2000), *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Leech, G., M. Hundt, C. Mair and N. Smith (2009), *Change in Contemporary English: A Grammatical Study*. Cambridge: CUP.

Schneider, E. (2007), *Postcolonial English: Varieties of English around the World*. Cambridge: CUP.

Wells, J. (2008), *Longman Pronunciation Dictionary*, 3rd edition. Harlow: Longman.

Part 1. Corpus development and corpus interrogation

An electronic corpus of *Letters of Artisans and the Labouring Poor* (England, c. 1750-1835): compilation principles and coding conventions

Anita Auer, Mikko Laitinen**, Moragh Gordon*** and Tony Fairman*****

*University of Lausanne
**Linnaeus University
***Utrecht University
****Independent Scholar

Abstract

This paper presents a collaborative project that focuses on letters of artisans and the labouring poor in England, c. 1750-1835 (LALP). The project's objective is to create a corpus that allows for new research perspectives regarding the diachronic development of the English language by adding data representing language of the lower classes. An opportunity for an insight into the language use of the labouring poor has been provided by the laws for poor relief, which permitted people in need to apply for out-relief from parish funds during the period 1795-1834. For the last 18 years, the independent scholar Tony Fairman has collected and transcribed more than 2000 poor relief application letters and other letters by artisans and the labouring poor. In this project Fairman's letter collection is being converted into an electronic corpus. Apart from converting the material into electronic form, the transcribed texts will be supplemented with contextual information and manuscript images. This paper presents the letter material, it describes the conversion of the letter collection into a corpus and discusses some of the problems and challenges in the conversion process.[1]

1. Introduction

One of the limitations of diachronic corpora is that they give us "at best, an inaccurate and skewed picture of spoken language", as pointed out by Rissanen (2008: 60). This limitation concerns in particular the field of socio-historical linguistics, which applies sociolinguistic methods (Weinreich, Labov and Herzog 1968, Labov 1972, 1994, 2001, Chambers 1995) to historical data with the aim of shedding light on the actuation and diffusion of linguistic changes in earlier stages of society. Considering that variationist sociolinguistic theory is by and large based on findings from spoken language, corpus linguists working in the field of socio-historical linguistics have aimed at putting together collections of written material that reflect spoken language as closely as possible. For English, for instance, selected digitised collections that have been compiled and made available to the research community are (a) letters in the *Corpus of Early English Correspondence* (Nevalainen and Raumolin-Brunberg 2003), (b) direct speech from witness depositions, didactic works, fiction and comedy in the *Corpus of*

English Dialogues 1560-1760 (CED) (Culpeper and Kytö 2010), and (c) the transcriptions of spoken English used in the *Old Bailey,* London's criminal court, from the seventeenth through to the early twentieth centuries (Huber 2007).

These records of speech-like texts are of great value, but, as Rissanen (2008) rightly points out, they are at best an imperfect rendering of speech in the past. After all, these texts may not always represent the spoken language of daily life; particularly because corpora containing speech-like texts often contain limited social variation as the informants mainly represent the top social layers, or they contain material recorded in unusual social situations such as the court room. Therefore, in recent years increasing efforts have been made in English corpus linguistics to unearth and digitise vernacular materials from below, that is to say from the lowest social layers (see selected articles in Auer, Schreier and Watts eds 2014). Many of these corpus projects focus on letters. Illustrative examples include for instance the *Corpus of Older African American Letters* (described in Siebers 2014), the *Hamburg Corpus of Irish English*, compiled by a group led by Peter Siemund and Lukas Pietsch at Hamburg University, the work on Scottish emigrant letters (Dossena 2012) and on the early nineteenth-century settler letters from the Cape Colony (Włodarczyk 2010, 2013).

Digitising written records produced by people from the lower layers helps to broaden the social scope of materials in diachronic linguistics. To date, the upper layers of society and men are over-represented in diachronic corpora. Considering that these materials have served as the basis for standard histories of the English language, it may be argued that this history is skewed towards the elite layers of society. This is in particular so if we consider the data representing the classically-educated people relative to their size in population. For instance at the beginning of the nineteenth century, England was a country with some nine to ten million inhabitants (Hilton 2006). In terms of income, the upper strata, whose income was over £800 per year and who received classical education, consisted of less than 100,000 people, and roughly one-third belonged to the middle order, earning over £10 per month. The great majority, some seven million people, earned less than ten pounds a month. These are the people who form the overwhelming majority; yet their voices, not to mention those at the very bottom of the social ladder (i.e. labouring poor), are to date rarely represented in diachronic corpora. Thus, corpus compilers have attempted to remedy social imbalances through the sampling process by giving preference to certain social groups, such as women and lower classes, or by collecting material from text types that represent these social groups. This practice is exemplified in Culpeper and Kytö (2010: 26) who point out that when collecting the CED, preference was given to under-represented groups like "women and/or the lower ranks" whose voices can for instance be heard in witness depositions dealing with witchcraft. Similar to Culpeper and Kytö (2010), we are concerned with making the voices of hitherto under-represented social groups heard.

In this article we introduce and discuss unique letter material written by the labouring poor in England during the late eighteenth and early nineteenth centuries. Our material consists of applications for out-relief and other correspondence, which have survived as records of the laws for parish relief for people in distress. This system provided help for people who had moved outside their home parishes; if they found themselves in dire straits, they were forced to write to their home parishes asking and arguing for help. Our informants were in many cases people who had not received much literacy training since elementary compulsory schooling (1st Education Act) was only introduced in 1870. The collected letter material thus provides data from the lowest social layer from the Late Modern English period (1700-1900).

In order to understand the social context of the collection *Letters of Artisans and the Labouring Poor* (LALP), one needs to be familiar with the poor relief system during the Late Modern English period. It was based on the Elizabethan Poor Laws of 1601, and this legislation had been amended in 1662 to formalise the notion of parish settlement so that each individual became chargeable to one parish at a time (Hitchcock, King and Sharpe 1997). The economic development in the eighteenth century, which was closely inter-related with the industrialisation and the enclosure of common and waste lands in rural areas, led to an increase in mobility all over England; this had a direct impact on the out-relief expenditure. Lindert (1997) estimates that the share of parish relief in the Gross Domestic Product in England increased from circa 1% in the mid-eighteenth century to nearly 3% by the mid-1830s. This increase was most likely linked to the Removal Act of 1795 which forced people to apply for out-relief by decreeing that the non-settled person who did so could not be removed, i.e. taken back to the parish of legal settlement. The act led to a massive increase in written documents from artisans and the labouring poor (seen for instance in the chronological distribution of the letters published in Sokoll 2001).

The out-relief system came to an end in 1834, but from a period of circa 40 years there exists plenty of letter material that is largely representative of the language of the lower classes in England. This material has survived in County Record Offices, from which the independent scholar Tony Fairman has collected and transcribed an illustrative sample over the past two decades (see for instance Fairman 2000, 2007). Our project, *Letters of Artisans and Labouring Poor* (England, c. 1750-1835), focuses on processing Fairman's transcriptions of over 2,000 letters and converting them into a searchable corpus with some 303,400 orthographic units. A first plain text version of the corpus, based on Tony Fairman's transcriptions, will be ready by the end of the year 2013. After that we are planning on creating an xml version of the corpus. As many of the transcriptions still require checking against the original sources that are held in archives and record offices, the plain text version of the corpus will not be made publicly available as long as revisions and corrections are being carried out. It is however possible to access the corpus at the universities of the corpus converters.

Apart from providing unique lower social layer vernacular material for historical sociolinguists, the project also aims at bridging a gap by working closely with historians and specialists in digital humanities. This interdisciplinary collaboration aims at enhancing usability and searchability of materials from the

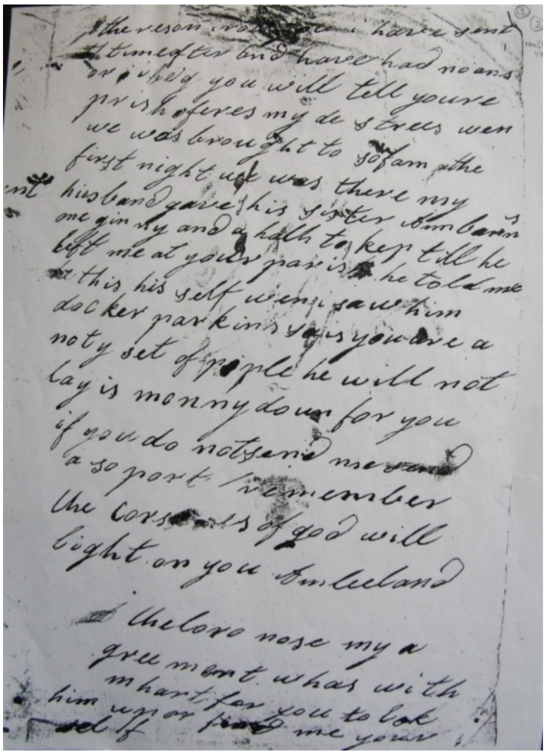
Late Modern period by exploring the possibilities to interconnect and merge additional resources, such as official records related to the correspondence.

2. The conversion of a letter collection into a corpus

The compilers of the CEEC, a state-of-the-art correspondence corpus with great value for the field of socio-historical English linguistics, have predominantly relied on published collections in which the editors have provided explanations of the transcription processes. Raumolin-Brunberg and Nevalainen (2007) point out that their work has included some degree of qualitative judgments related to the reliability of these processes. In our case, the starting point has been Fairman's transcriptions, consisting of over 2,000 letters. When he started collecting and transcribing pauper letters and related correspondence, his aim was to capture as many of the physical properties of handwriting and the social backgrounds of the applicants as possible. Fairman's transcriptions allowed us to have plain text versions that conveyed textual as well as extra-textual features of the original manuscripts. However, since the transcriptions were made across a time-span of two decades and due to the idiosyncratic nature of the material, inconsistencies with regard to transcription practices and coding were inevitable. For instance, Fairman's transcription practices have developed over the years and therefore also changed quite significantly. Whenever possible, Fairman's transcriptions were checked against the originals. In some cases our starting point was a photocopy of the original documents accompanied by the transcription of the letter; in most of the cases, however, we had to start with existing transcriptions but we are still planning on checking them against the originals held in the record offices.

Considering that Fairman's original aim was to investigate lower class writing rather than to compile a searchable electronic corpus, his transcriptions and the coding used were not necessarily suitable to be transferred to plain text versions. One of the major steps in the conversion process was thus to render Fairman's codings into a digital format. In this process we could rely on the conventions that have been established by the Helsinki corpus team (Kytö 1996) and other compilers of letter corpora (e.g. van Bergen and Denison 2007). As our data differs from other correspondence data in that it is highly idiosyncratic, i.e. in particular with respect to spelling and punctuation, some adjustments and additions were needed in order to deal with the heterogeneity of the material. Since it was our goal to create a corpus that could serve an interdisciplinary research community, we have tried to maintain as much of the richness of the material as possible. This is of course also where the challenge lies because each research discipline requires its own set of data and also focuses on different features of the original documents. For instance, a social historian might attach more importance to the contents of a letter and the social background of its writer; a palaeographer is primarily interested in the handwriting of a letter, whereas a historical sociolinguist will be interested in, for instance, lexical or grammatical items and/or the authors' social backgrounds. Considering that it is a challenging task to facilitate all these different research purposes, a plain text version alone might not be sufficient to capture the richness of the material. Our ultimate aim is therefore to have, next to a plain text version, a normalised version that makes searching for lexico-grammatical items easier and also enables POS-tagging, as well as a multi-media version which allows for searches for lexical elements and grammatical structures and for access to the image of the original at the same time.

Figure 1 below shows a photocopy of an original letter followed by its original transcription. The most striking feature is the idiosyncratic spelling. The letter also contains some self-corrections. As can be seen in the transcription following the letter, all of these features as well as the original format are preserved in Fairman's transcription, together with information about the holding archive. In selected letters Fairman even supplied information with regard to the way in which the author formed or deleted graphs. Although the transcription represents the original well, this version is not easily searchable using standard concordance interfaces.



the reson [illegible] have sent
time after and have had no ans
or i hop you will tell youre
prish ofeces my de strees wen
we was brought to sofam the
first night we was there my
husband gave his sister [illegible]
one ginny and a halth to kep till he
left me at your paris he told me
this his self wen i saw him
docker parkins says you are a
noty set of pepple he will not
lay is monny down for you
if you do not send me
a soport remember
the corses of god will
light on you [illegible]

the lord nose my a
gree ment whas with
m hart for you to lok
him upor [illegible] me your
self

Figure 1. Photocopy of letter from Swaffham Prior

CA/SW/7: (My No. 12+)
Swaffham Prior, P150/18/1

the reson i ro[te y]ou i have sent
~~th~~ time after and have had no ans
or i beg you will tell youre
prish ofcres my de strees wen
we was brought to sofam the
first night we was there my
husband gave his sister Ann barns
one ginny and a halh to kep till he [sic]
left me at your parish he told me
~~he~~ this his self wen i saw him
docker parkins sais you are a
noty set of pXple he will not [X = blot]
lay is monny down for you
if you do not send me ~~send~~
a soport / remember
the cors[BLOTS}es of god will
light on you Am leeland [sic: = Ann]
~~it~~ the lord nose my a
greement whas with
m hart for you to lok
him up or find me your
self

As far as the transcription and coding are concerned, one of the project's main aims was to provide diplomatic transcriptions of the original manuscripts, preserving the original spelling, lineation, punctuation, insertions, interlineations, blots and strike-throughs. The transcription below illustrates the revised transcription of the original letter (Figure 1 above). As can be seen, this transcription differs slightly from Fairman's transcription. The lineation has been preserved, whereas the emendations in lines 2, 10, 14 and 18 and the blots in lines 12 and 16 are pointed out in comments and via codes.

1 the reson i ro{te y}ou i have sent
th [^th CROSSED OUT^] time after and have had no ans
or i beg you will tell youre
prish ofcres my de strees wen
5 we was brought to sofam the
first night we was there my
husband gave his sister Ann barns
one ginny and a halh to kep till he
left me at your parish he told me
10 he [^he CROSSED OUT^] this his self wen i saw him

docker parkins sais you are a
noty set of p[^BLOT^]ple he will not
lay is monny down for you
if you do not send me [^send CROSSED OUT^]
a soport / remember
the cors[^BLOT^]es of god will
light on you Am leeland
it [^it CROSSED OUT^] the lord nose my a
greement whas with
m hart for you to lok
him up or find me your
self

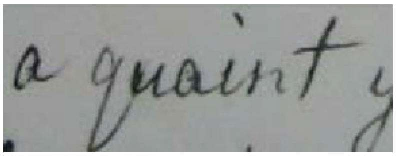
2.1 Text-level coding

The discussion in this section illustrates our transcription processes and coding schemes. The transcription codes are based on those used in the *Helsinki Corpus of English Texts* (Kytö 1996) and were modified for our material. Table 1 below shows the codes that were used to transcribe general textual features of the original manuscripts, such as punctuation, superscript and underlining. Table 2 below shows the codes that were used to express paratextual features, such as strike-throughs and interlineations.

One of the major challenges in the transcription process was that we had to deal with material that was often written by highly inexperienced writers with restricted writing skills. As a consequence it was sometimes difficult for the transcribers to decipher a graph or even whole words. In some cases it was also hard to determine whether an author used upper or lower case. Where ambiguity could not be ruled out, a code was used, as adopted from van Bergen and Denison (2007), which indicates conjectural readings. We did however not apply this for uncertain cases of capitalisation since that would negatively affect the readability of the texts (a problem that also has been pointed out in van Bergen and Denison 2007: 8). As regards our transcription practice, the following points were observed in the conversion process: The line division of the original manuscripts was generally preserved. The indentations and large spaces of the source texts were imitated as closely as possible by using interspacing and word processor format indentations. This implies that the indentations and spaces in the transcription are an approximation and not an exact copy of the indentions and spaces as found in the source text.

Page breaks were indicated by [^TP^] (turn page) when the letter continued on the reverse of the paper or, when it continued on a new page [^NS^] (new sheet). Blank lines were indicated by ‘our comment’.

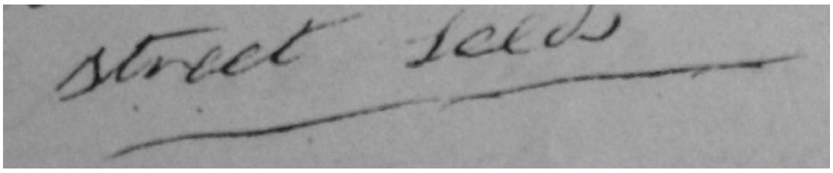
In the plain text version, the original word boundaries were maintained. This included word divisions across line breaks. Although this greatly hampers searchability, we also tried to preserve word joinings and divisions that deviated from modern standard spelling. Because the spelling generally varies greatly as well, the plain text version is less suitable for word searches. The normalised version will be more useful for this purpose. Figure 2 below illustrates how the author divided *acquaint* (here spelled *a quaint*) into two separate units. This is just one example of the many different ways in which the authors of the corpus either joined or separated units. In fact, it was often a matter of speculation of whether there was a word boundary or not. In the example below the *a* is clearly separated from the rest of the word but in many cases the boundaries were rather fuzzy.



a quaint y

Figure 2. (Addingham, 49D90/6/b21, Yorkshire)

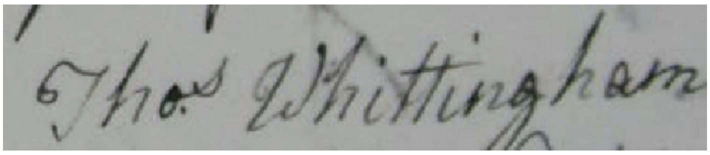
If a word was underlined in the original manuscript, it was followed by ‘our comment’ stating [^UNDERLINED^] in the transcription. If more than one word was underlined, the words were also included in the comment: xx x [^xx x UNDERLINED^]. When several words, sentences or lines were underlined, the first underlined word and the last underlined word were contained in the comment: [^xx…xx UNDERLINED^]. The example below, Figure 3, was transcribed as follows: street Leeds [^street Leeds UNDERLINED^]



street Leeds

Figure 3. (Beverly, St. Mary’s, PE/1/723, Yorkshire)

Items written in superscript in the manuscripts were expressed with the = sign. Words such as Si^{r} or abbreviations such as 20th in the source text were converted into Si=r= and 20=th= in the transcription. Figure 4 below is an abbreviation of the name *Thomas* with a superscript *s*, which has been transcribed as follows: Tho.=s= Whittingham



Tho^{s} Whittingham

Figure 4. (Addingham, 49D90/6/b21, Yorkshire)

Items written in subscript in the manuscripts have been expressed with the = sign followed by our comment stating [^SUBSCRIPT^]. In Figure 5 below, the last two graphs were crammed into the page and written in subscript position: my Childr=en=[^SUBSCRIPT^]. It needs to be pointed out that interpreting these kinds of features are subject to speculation and in our case it was sometimes difficult to define and label features such as subscript writings unambiguously.

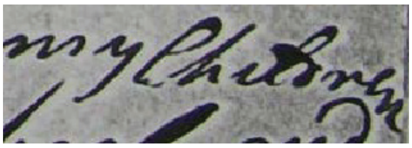


Figure 5. (Kirkby Lonsdale, WPR/19/1919/35, Westmoreland)

Table 1. General format codes

General format	**Codes**
Page Breaks	[^NP^] [^NS^]
Punctuation marks	! , .- / : ; ? £ () “” =
Underlining	[^UNDERLINED^]
Superscript	yyy=x=
Subscript	yyy=x=[^SUBSCRIPT ^]

In addition to the general format of the transcription, there are also codes indicating the paratextual features of the manuscripts. The following transcription codification conventions were used:

Table 2. Text-level coding

Text-level coding	**Codes**
Our Comment	[^ yy XXXX^]
Conjectural Reading	{xxx}
Illegible material with approximation of number of illegible characters	{***}
Illegible material without an indication of the numbers of characters	{*…}
Deleted Items	[^xxx CROSSED OUT^]
Deleted items with some parts uncertain	[^yyy {xx} yy CROSSED OUT^]
Deleted items with some parts illegible	[^xxx{**}xx CROSSED OUT^]
Interlineations and insertions	yy=xx=[^INTERLINEATION^]/[^INSERTION^]

Any information that was deemed relevant was incorporated in ‘our comment.’ These comments were typed in capitals, and any letters or words cited from the transcription were typed in normal font. For instance, when a letter contained an annotation in another hand, this was indicated by a comment followed by the transcribed annotation: [^ANOTHER HAND^] xxxxxxxx. Comments were also used to indicate any feature that could not be expressed by a code, for instance, in Figure 6 below *and* is overwritten by the phrase *I get*. This was represented as follows: [^I get OVERWRITES and^].

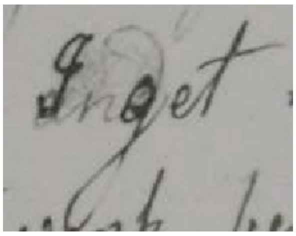


Figure 6. (Addingham/49D90/6/b/21, Yorkshire)

In some instances, it was necessary to explicitly point out when a reading was uncertain. This was indicated with curly brackets around individual letters or larger units. For instance, {k}ing indicates where the transcriber had difficulties determining the actual shape of one character, or {king} when a whole unit of characters was hard to define. We also marked conjectural reading when graphs could not be read because of damaged paper or faded ink. In the example below, Figure 7, the word was partially illegible because the paper was torn. As this word was written on the date line, the context revealed that the word was most likely *March*. However, the *h* was no longer legible and this graph was thus enclosed by curly brackets: Marc{h} [^TEAR^].

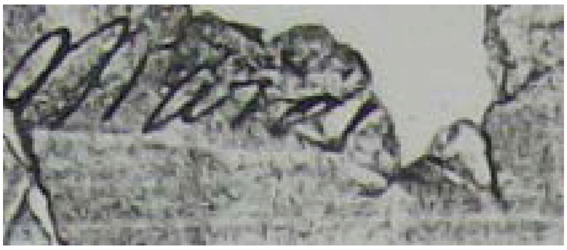
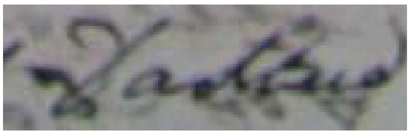


Figure 7. (Harthill All Hallows, PR47/91/4, Yorkshire)

Illegible material was represented by an asterisk for each illegible character enclosed by curly brackets. For instance, {***} indicates that there were three illegible characters, or an approximation of it. If it was not possible to estimate the number of illegible characters, {*…} was used. Sometimes a unit was only partly illegible. In such a case, it was presented with: {*}x{**}. Figure

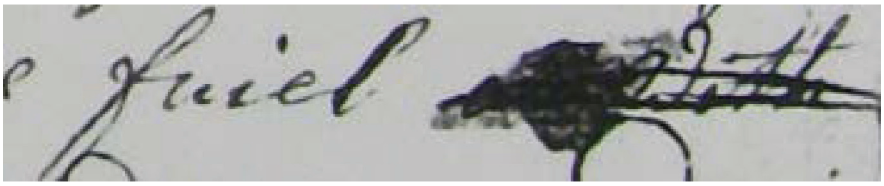
8 below illustrates a case in which the transcriber could only decipher the first two graphs, the others were unclear. There appeared to be three characters but, even within the context of the sentence, no sense could be made of them. The word was therefore transcribed as follows: da{***}.



da

Figure 8. (Kirkby Lonsdale, WPR/19/1825/5, Westmorland)

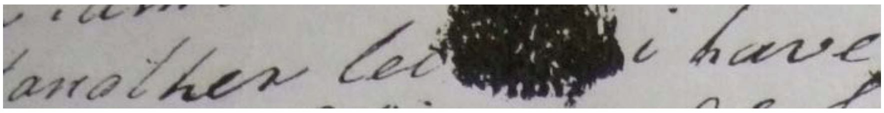
When the author of the original manuscript had deleted an item by crossing or rubbing it out, as in Figure 9, this was indicated with a comment containing a reconstruction of the word that was crossed out: [^xxx CROSSED OUT^]: Fuiel with [^with CROSSED OUT^].



Fuiel ~~with~~

Figure 9. (Harthill All Hallows, PR47/91/4, Yorkshire)

When possible, a comment was added to clarify the cause of the illegibility. For instance, when a word was crossed out to such a degree that some characters were no longer or hardly legible, a combination of two text level codes was used: [^yy{xx} yy CROSSED OUT^] and [^xx{***}xx CROSSED OUT^]. In Figure 10 below only the first three graphs were legible, and the rest was blotted out: another let {*…} [^BLOT^] I have.



another let i have

Figure 10. (Swaffham Prior, P150/18/1, Cambridge)

When the author inserted writings between lines or within words, the following code was used: yy=xxxx= [^INTERLINEATION^]. When a letter was inserted: yy=xxx= [^INSERTION^] was used. In the example below *you* was inserted: to =you= [^INSERTION^] I hope in.

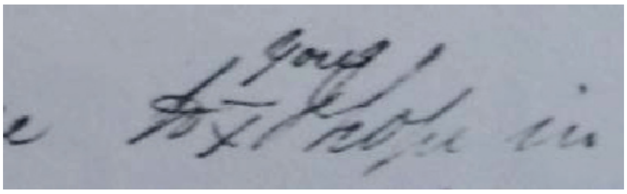


Figure 11. (Brampton 45/18/2, Huntingdonshire)

2.2 Challenges in the normalisation process

As mentioned above in the previous section, the plain text version maintains the original spelling and word boundaries. However, the highly idiosyncratic nature of both spelling and word boundaries poses a problem for concordance tools and other search interfaces, and one solution which we are testing is to normalise the spelling. It should be pointed out that while a version with normalised spelling facilitates searching the material, it would naturally not be of great use for studying lexico-grammatical variability. The aim is therefore to produce a second version of the corpus with in-text normalisation tags which make it possible to search for a lexeme using its standard spelling but having access to the variant spelling forms at the same time.

To explore the possibilities for normalising spelling in a heterogeneous vernacular material, we have used VARD, a variation detector tool, developed by Baron and Rayson (2009), that can detect, normalise and tag variant spellings. Below is an example of an in-text tag, which shows that the author used *goon* for the past participle *gone*. The first set of angled brackets forms the start tag and encloses the original word that has been normalised, followed by the normalised word; the latter is not enclosed by brackets so that a concordance programme or any other search programme can retrieve *gone*. The second set of angled brackets form the end tag.

have been <normalised orig="goon" auto=false">gone</normalised> this

According to Baron and Rayson (2009), the VARD tool was initially developed to deal with Early Modern English spelling variation and could therefore only detect a limited amount of variations in highly idiosyncratic correspondence material from lower social layers. The updated version is more advanced and “employs techniques from modern spell checking software to search for potential variants and find candidate equivalents for variants found” (Baron and Rayson 2009). One of the advantages of the updated version is that it uses ‘phonetic matching techniques’ that allow for easier detection of speech reflection in spelling. Moreover, the accurate equivalent standard spelling can be provided. In the case of the LALP corpus, this feature is very helpful as the variant spelling often seems to reflect pronunciation. Another advantage is that the programme can be trained to recognise certain variants and to supply corresponding variants. However, this may only be of benefit when the amount of variation is limited and to a certain extent predictable. One problem with these vernacular letters is that the spelling often varies in highly unpredictable ways, sometimes per author or even per letter, and a fully automatic normalisation will probably not give a satisfactory result. Nonetheless, the programme can still be of some benefit since it detects most variants and allows the corpus compiler to go through the detected words manually. It also enables choosing a normalised equivalent from a list of close candidates, and if one is not present, it is possible to enter the lexeme manually. Considering that the programme automatically provides in-text normalisation tags, the amount of manual labour is reduced.

What is still challenging is that the letter writers often separated words that are joined according to modern standard spelling, as also pointed out above. For instance, *someone* may be spelled as *some* and *one*, or *another* as *a nother.* None of these words would be detected as variants since they, with the exception of *nother*, correspond with the modern spelling variants. This is also the case with words that correspond with modern standard variants but differ in meaning from words used in the particular context, as is the case in extracts (1-5) below:

(1) I **ham** very sory to make this Apell
(2) my Little Boy **his** orderd
(3) i **most** have sent before
(4) he is and **as** been pain nine
(5) Send my Girls money wich is but one Quarter **dew** to me

While VARD allows speeding up the normalisation process, manual checking by the corpus compiler is inevitable, in particular with respect to the cases listed above.

3. Providing contextual information in a header

Social and contextual information that enables searching for letter material has been provided in a set of header lines attached to each individual letter. These header lines are made up of two angled brackets that enclose a code and a corresponding value, which is illustrated below. An 'x' after the code was used to indicate that a value was not applicable or available.

<F File name>
<E Photocopy available: 1/0>
<Q Identification number archive>
<B Name author>
<H Age author>
<G Sex author>
<R Place author>
<N Name applicant>

<O Other applicants>
<A Age applicant>
<S Sex applicant>
<P Place applicant>
<L Legal parish of settlement>
<D Date>
<X Number of letters>
<W Number of words>
<C Topic: Application/Related correspondence>
<M Miscellaneous>

<F> File name

The file name was made up of the date of writing, the surname of the author and the county where the original manuscript was found. e.g. _1781_03_30_Johnson_5 (_year_month_day_surname_countynumber). The county number corresponds with the county numbering in Williams (2004: vi). In case one of the values was not known, as pointed out above, an x was used to indicate this. To avoid filename clashes, numbers were added. For instance, there were two letters from Lancashire whose dates and authors were unknown, which meant that the file names only contained xs and a county number, e.g. _xxxx_xx_xx_x_3. To avoid double file names the first anonymous file lacking a date was labelled anonymous 1 and the second anonymous 2, which may be illustrated by _xxxx_xx_xx_anonymous1_3. Similarly, when an author had written more than one letter without any indication of a date, a number was added to the author's name: _xxxx_xx_xx_Johnson1_3

<E> Availability of Photocopy

As pointed out earlier, for some of the letters digital images of the original manuscript are available. When a photocopy was available this was indicated with 1; and with zero when no photocopy was available.

<Q> Identification Number Archive

This header provides information about where the original manuscript is stored and how it is registered in the archive.

<B> Name Author

We made a distinction between the author of the letter and the applicant. If known, the name of the author was provided, i.e. first name (in full if possible), followed by surname. As is often the case in socio-historical linguistics, caution must be applied when it comes to genuine authorship (Hernández-Campoy and Schilling 2012). In our material, such cases include instances in which (a) several letters in the collection were signed with one name but were written in different hands, or (b) letters started with a reference to the name of the applicant, who was then consistently referred to as *him* or *her* but then the letter was signed with the applicant's name, as the following example illustrates (Wroughton: 551/96, Wiltshire):

> **Mariam King** would be very mutch
> Oblight to you for to Send **her** her
> Money as **She** is a Live & in Great Want
> [...]
> so no more at presant
> **Mariam King**

The way in which Mariam King is mentioned strongly suggests that someone was writing the letter of application on behalf of her, even though it was signed with her name. (c) In some cases a wife wrote on behalf of her husband and signed with her husband's name. (d) Selected letters in the collection were signed by more than one author, or the author used a formula that contained the name of his/her spouse as well, all of which made it challenging to determine who had in fact written the letter.

When authorship was doubtful due to the factors listed above, this was indicated in the 'Miscellaneous' section. Furthermore, the spelling of names often varied. For the sake of consistency, the most frequently used variant was picked to be used in the file names and in the <B> section. The other variants were listed in the 'Miscellaneous' section.

<H> Age Author

If the age and the year of birth of the author were known this was indicated as follows: Age/Year of birth.

<G> Sex Author

If known, the sex of the author was indicated.

<R> Place author

This is the place, as indicated by the author, from which the letter was sent. As the place name was often spelled in a non-standard way, it was at times difficult to determine which place name was meant. In case of uncertainty, this was indicated with a question mark behind the assumed place name. The place name as originally spelled by the author was then listed in the 'Miscellaneous' section.

<N> Name Applicant

In this section the name of the person who applied for relief, or the name of someone who could be connected to an application for relief in some other way, or the name of a parish was provided. In case that the author was also the applicant, the same information was given in both the author and the applicant line. As for variant spelling of the name, the same approach as used in line <B> was taken.

<O> Other Applicants

If the application concerned more than one applicant, this was indicated, and all name(s) were provided if available.

<A> Age Applicant

If known, the age and the year of birth of the applicant were indicated like this: Age/Year of birth.

<S> Sex Applicant

This information was provided if known.

<P> Place Applicant

In case that the applicant and the author were the same person, the place of writing as indicated by the author was used (and assumed that it was the place of residence). If the applicant was not the author, it was more difficult to infer the place of residence from the letter. Due to non-standard spelling of the place names, it was sometimes difficult to determine which place name was intended. In case of uncertainty, the assumed place name was marked with a question mark. The place name as spelled by the author was listed in the 'Miscellaneous' section.

<L> Parish of Legal Settlement

The legal place of settlement, i.e. the 'home parish', is where an applicant had his or her settlement rights and where he or she could apply for relief. Settlement rights could for instance be established by birth, marriage, apprenticeship or by renting a property in that parish for a certain length of time (cf. White 2004: 280, for more details see Auer and Fairman 2013). The notion of legal settlement implied that the parish was responsible for providing out-relief for an applicant residing elsewhere.

<D> Date

Writers usually noted dates in their applications. In cases where no date or year was noted, this information could often be inferred from the postage stamp. Sometimes a date was given but annotated in another hand, and in such cases the year or date was followed by a question mark in the header line. In some cases where no date was given, there was an indication in the letter that it was written in a certain period. A letter estimated to have been written between 1810 and 1820 was noted as 1810-20 on the date line. The filename indicates this with adding an 's' to the year that spans the decade(s) in which a letter must have been written. Lastly, sometimes, when there was no date, an approximation of the earliest possible date was provided on the basis of contextual information, noted down as a date followed by + to indicate that it was an estimation.

<X> Number of Letters

This header line provides information about the number of letters (contained in the corpus) from one author. It needs to be pointed out that the number of letters reflects the number of times a letter was signed with the same signature. When genuine authorship remained doubtful, this was indicated in the 'Miscellaneous' section.

<W> Number of Words
This line provides an indication of the number of words written by the author. The aim was to count all the lexemes written by one author, including words that were crossed out, salutations, valedictions, and addresses. A word that was divided into two words, or two words that were joined into one word, were counted as either two units or as one unit respectively. A word that was broken up across line breaks was also counted as two words. Annotations in other hands and our own comments were excluded from the word count.

<C> Topic
The topics of these letters were divided into two categories. A letter was either an application, or it was related to the application process in some way. The values used were therefore either 'application' or 'related correspondence'. An application was usually written/initiated by the applicant and addressed to the overseer. However, sometimes a letter was addressed to a relative who was asked to apply on behalf of the author. Although this may be considered an indirect application, we categorised such a letter as application. The term 'related correspondence' comprises correspondence concerning information about marriage certificates, apprenticeships, reports of abuse or fraud regarding out-relief, or correspondence between overseers of the home parish and the out-parish of an applicant.

<M> Miscellaneous
This section was reserved for remarks that we deemed relevant, e.g. doubtful authorship, questionable place names, variant name spelling and dates. Moreover, in some cases letters by different authors were related to each other in some way. The names of the related files were listed in this line.

4. Outlook and possibilities

This collection of letters of artisans and labouring poor adds a new layer to existing diachronic correspondence data. The new data of a hitherto under-represented layer of society are of great value because socio-historical linguists now have the possibility to compare language use in the Late Modern English period across all social levels. This will allow for more accurate descriptions of linguistic variability in Late Modern England. As these letters by and large originate from people who have received little schooling in comparison to the upper layers of society, a number of questions may be raised. For instance, did the lower classes use more informal styles in a formal and official communicative situation? The material also allows us to test the question as to what extent the written language of artisans and the labouring poor reflects dialect use and/or spoken language (cf. Fairman 2006). In addition, the Late Modern English period was the time during which accent developed as a social symbol and numerous pronunciation dictionaries and elocution guides aimed at lower and middle sections of the society were published (see Mugglestone 2003 for an extensive discussion). It is clear from these elocution guides that a social meaning was given to sensitive variables such as dropped initial ‘h’ and word ending ‘g’. Comparing between given pronunciation advice as reflected in these manuals and speech reflections in the LALP corpus may shed new light on norms and actual practices at the time. This information based on actual speech could thus supplement the evidence contained in eighteenth-century pronunciation dictionaries (see Beal 1999: ch. 3). Apart from speech reflection, these letters may also shed light on diachronic variation in dialects (cf. present-day synchronic descriptions in Trudgill 1999, Szmrecsanyi 2012). After all, these letters are records of mobility in which two types of background information are nearly always known. These consist of the information on the legal parish of settlement (i.e. the places where the letters were sent to and where they had at some point established residenceship) and the information from where people were applying for out-relief, that is to say the locations where they currently resided. As our discussion above (Section 3) shows, the material can be searched according to these regional parameters.

In addition, sending these applications meant crossing physical and social spaces, which in many instances must have been a new way of communicating for many individuals. These letters are authentic accounts of how people had to make use of written vernaculars to negotiate with the established societal system, and the applicants had to engage in a process of socialisation of developing one form of genre literacy (cf. Taavitsainen 2010 on the evolution of genres in medical writing). It will therefore be useful to draw from the theoretical insights of scholars who have explored similar types of socially uneven communication in other contexts. These include the theoretical insights in sociolinguistics of globalisation (Blommaert 2010). This approach sees that access to forms and their normative control is always unevenly distributed. Language use always has a normative dimension, as some forms and pragmatic solutions are more valuable than others depending on the social situation anchored in time and space where they occur (cf. present-day asylum seeker discourses in Maryns and Blommaert 2001). So far, the entire enterprise of sociolinguistics of globalisation has been seen to be associated with present-day societies solely but this material clearly shows how the framework needs to be tested in time (cf. Laitinen 2014).

Lastly, the material presented here, though it is unique in English corpus linguistics, forms one part of a larger set of letters that have emerged as materials in diachronic linguistics recently (cf. Montgomery, Fuller and DeMarse 1993, Schneider and Montgomery 2001 on the work carried out on the formation of American English). These materials “from below the social spectrum” consist of emigrant letters from around the world, and are not only restricted to letters written in English. These letters are located in archives around the world, and we aim at combining the expertise of (a) historians in locating and providing the historical context, (b) linguists for transcribing and analysing the language of daily life, and (c) specialists in digital humanities for contributing to digitising the material, and improving interconnectivity between the various digital collections. Our broad objective is to improve access to digital resources that are of interest not only to academics but also to the general public. Our main project result is to make the letters available in a format that ensures easy access to the material and its preservation for future generations.

Notes

[1] We wish to thank the two anonymous peer reviewers for their valuable comments. The authors are responsible for any remaining errors.

References

Auer, A. and T. Fairman (2013), 'Letters of artisans and the labouring poor (England, *c.* 1750-1835)', in: P. Bennett, M. Durrell, S. Scheible and R. J. Whitt (eds) *New Methods in Historical Corpora*. Narr: Tübingen. 77-91.

Auer, A., D. Schreier and R. J. Watts (eds) (2014), *Letter Writing and Language Change*. Cambridge: CUP.

Baron, A. and P. Rayson (2009), 'Automatic standardization of texts containing spelling variation, how much training data do you need?', in: M. Mahlberg, V. González-Díaz and C. Smith (eds) *Proceedings of the Corpus Linguistics Conference, CL2009, University of Liverpool, UK, 20-23 July 2009.* Available online at http://ucrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf.

Beal, J. C. (1999), *English Pronunciation in the Eighteenth Century. Thomas Spence's 'Grand Repository of the English Language'*. Oxford: OUP.

Blommaert, J. (2010), *The Sociolinguistics of Globalization.* Cambridge: CUP.

Chambers, J.K. (1995), *Sociolinguistic Theory*. Oxford: Blackwell.

Culpeper, J. and M. Kytö (2010), *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: CUP.

Dossena, M (2012), '"I write you these few lines": Metacommunication and pragmatics in nineteenth-century Scottish emigrants' letters', in: U. Busse and A. Hübler (eds) *Investigations into the Meta-Communicative Lexicon of English: A Contribution to Historical Pragmatics*. Amsterdam: Benjamins. 45-63.

Fairman, T. (2000), 'English pauper letters 1830-1834, and the English language', in: D. Barton and N. Hall (eds) *Letter Writing as a Social Practice*. Amsterdam: Benjamins. 63-82.

Fairman, T. (2006), ‘Words in English record office documents of the early 1800s’, in: M. Kytö, M. Rydén and E. Smitterberg (eds) *Nineteenth-Century English: Stability and Change*. Cambridge: CUP. 56-88.

Fairman, T. (2007), ‘Writing and “the Standard”: England, 1795’, *Multilingua* 2: 167-201.

Hernández-Campoy, J. M. and N. Schilling (2012), ‘The application of the quantitative paradigm to historical sociolinguistics: problems with generalizibility principle’, in: J. M. Hernandez-Campoy and J. C. Conde-Silvestre (eds) *The Handbook of Historical Sociolinguistics*. London: Blackwell-Wiley. 104-120.

Hilton, B. (2006), *A Mad, Bad, and Dangerous People? England: 1783-1846*. Oxford: Clarendon Press.

Hitchcock, T., P. King and P. Sharpe (eds) (1997), *Chronicling Poverty. The Voices and Strategies of the English Poor, 1640-1840*. London: Macmillan.

Huber, M. (2007), ‘The Old Bailey proceedings, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English’, in: A. Meurman-Solin and A. Nurmi (eds) *Annotating Variation and Change* (Studies in Variation, Contacts and Change in English 1). Available online at http://www.helsinki.fi/varieng/journal/volumes/01/huber/ (last accessed on June 7, 2011).

Kytö, M. (1996), *Manual to the Diachronic Part of the Helsinki Corpus of English Texts. Coding Conventions and Lists of Source Texts*. 3rd ed. Helsinki: University of Helsinki, Department of English.

Labov, W. (1972), *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1994), *Principles of Linguistic Change. Volume 1: Internal Factors*. Oxford: Blackwell.

Labov, W. (2001), *Principles of Linguistic Change. Volume 2: External Factors*. Oxford: Blackwell.

Laitinen, M. (2014), ‘Early nineteenth-century pauper letters’, in: A. Auer, D. Schreier and R. J. Watts (eds) *Letter Writing and Language Change*. Cambridge: CUP.

Lindert, P. H. (1997), ‘Unequal living standards’, in: R. Floud and D. McCloskey (eds) *The Economic History of Britain since 1700*. Cambridge: CUP. 357-386.

Maryns, K. and J. Blommaert (2001), ‘Stylistic and thematic shifting as a narrative resource: assessing asylum seekers’ repertoires’, *Multilingua* 20: 61-82.

Montgomery, M., J. M. Fuller and S. DeMarse (1993), ‘“The black men has wives and Sweet harts [and third person plural *-s*] Jest like the white men”: evidence for verbal *-s* from written documents on 19th-century African American speech’, *Language Variation and Change* 5: 335-357.

Mugglestone, L. (2003), “*Talking Proper”: The Rise of Accent as Social Symbol*. Oxford: OUP.

Nevalainen, T. and H. Raumolin-Brunberg (2003), *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London and New York: Pearson.

Raumolin-Brunberg, H. and T. Nevalainen (2007), 'Historical sociolinguistics: the corpus of Early English correspondence', in: J. C. Beal, K. P. Corrigan and H. L. Moisl (eds) *Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases*. Houndsmills: Palgrave Macmillan. 148-171.

Rissanen, M. (2008), 'Corpus linguistics and historical linguistics', in: A. Lüdeling and M. Kytö (eds) *Corpus Linguistics: An International Handbook* Vol 1. Berlin: Walter de Gruyter. 53-68.

Schneider, E. W. and M. Montgomery (2001), 'On the trail of early nonstandard grammar: An electronic Corpus of Southern U.S. Antebellum Overseers' Letters', *American Speech* 76: 388-409.

Siebers, L. (2014), 'Assessing heterogeneity', in: A. Auer, D. Schreier and R. J. Watts (eds) *Letter Writing and Language Change*. Cambridge: CUP.

Sokoll, T. (2001), *Essex Pauper Letters 1731-1837*. Oxford: OUP.

Szmrecsanyi, B. (2012), *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: CUP.

Taavitsainen, I. (2010), 'Discourse and genre dynamics in Early Modern English medical writing', in: I. Taavitsainen and P. Pahta (eds) *Early Modern English Medical Texts: Corpus Description and Studies*. Amsterdam: Benjamins.

Trudgill, P. (1999), *The Dialects of England*. Oxford: Blackwell.

van Bergen, L. and D. Denison (2007), 'A corpus of late eighteenth-century prose', in: J. C. Beal, K. P. Corrigan and H. L. Moisl (eds) *Creating and Digitizing Language Corpora. Vol. 2: Diachronic databases*. Basingstoke: Palgrave. 228-246.

Weinreich, U., W. Labov and M. Herzog (1968), 'Empirical foundations for a theory of language change', in: W. P. Lehmann and Y. Malkiel (eds) *Directions for Historical Linguistics*. Austin: University of Texas Press. 95-188.

Whyte, I. (2004), 'Migration and settlement', in: C. Williams (ed) *A Companion to Nineteenth-Century Britain*. Malden, MA: Blackwell. 273-286.

Williams, C. (2004), *A Companion to Nineteenth-Century Britain*. Malden, MA: Blackwell.

Włodarczyk, M (2010), 'Infinitives in the 1820 Settler letters of denunciation: what can a contextualized application of corpus-based results tell us about the expression of persuasion?' *Poznań Studies in Contemporary Linguistics* 46: 533-564.

Włodarczyk, M. (2013), '1820 Settler petitions in the Cape Colony: Genre dynamics and materiality.' *Journal of Historical Pragmatics* 14: 45-69.

Towards a corpus of eighteenth-century English phonology

Joan C. Beal and Ranjan Sen

University of Sheffield

Abstract

This paper gives an account of plans for constructing a searchable database of eighteenth-century English phonology. The project incorporates data from pronouncing dictionaries and other texts dealing with pronunciation published in the second half of the 18th century. The data will be recorded in the form of Unicode transcriptions of as many of the approximately 1,700 words used to exemplify John Wells' (1982) Standard Lexical Sets as appear in the eighteenth-century texts. Although all the eighteenth-century texts purported to describe the 'best' English, they were compiled by authors from different parts of the English-speaking world (mainly different regions of England, Scotland and Ireland but including some from North America) and so can provide evidence for geographical diffusion of innovations. (Beal 1999, C. Jones 2006). This paper provides an account of the design of this database and presents the results of a pilot study demonstrating how such a database can be used to answer questions concerning the chronological, social, geographical and phonological distribution of variation between /hw/ ~/w/ ~ /h/ in WHICH, WHO, NOWHERE, etc. which is of interest to sociolinguists, dialectologists and historical phonologists.

1. Introduction

The 'corpus revolution' has transformed the study of English historical linguistics, but, until relatively recently, historical corpora of English have tended to be compiled from Middle and Early Modern English materials, leaving the eighteenth and nineteenth centuries as the 'Cinderellas of English historical linguistic study' (C. Jones 1989: 272). Describing the then newly-compiled *Corpus of Late Modern English Texts* (CLMET), De Smet makes the following comment:

> Symptomatic of a certain neglect of anything beyond the 17th century is the fact that the *Helsinki Corpus*, until now the most important electronic corpus for the study of the history of English, takes its final cut-off point in 1710. (De Smet 2005: 69).

Although the *Helsinki Corpus* (Rissanen et al. 1991) does indeed stop at 1710, there are now several corpora of English texts from the eighteenth and/ or nineteenth centuries. The *Penn Parsed Corpus of Modern British English* (PPCMBE), released in 2010, takes up where the *Helsinki Corpus* left off and covers the period 1700-1914, whilst the *Corpus of Historical American English* (COHA) includes nineteenth-century American English texts and the *Corpus of*

Oz Early English (COOEE) (Fritz 2007) is compiled from English texts written in Australia, New Zealand and Norfolk Island between 1788 and 1900. ARCHER covers the period 1650-1990 and includes material from nine genres and both British and American English. Other Corpora, such as the *Corpus of Early English Correspondence Extension* (CEECE), the *Network of Eighteenth-century English Texts* (NEET) (Fitzmaurice 2007) and the Corpus of late Eighteenth-century Prose, concentrate on letters, whilst the *Old Bailey Corpus* has been compiled from the court documents originally digitised for the *Old Bailey Online* project. In addition to these corpora, scholars can now access electronic databases such as *Eighteenth-Century Collections Online* (ECCO), the *Eighteenth-Century English Grammars* (ECEG) database, the Chadwyck-Healey databases of eighteenth- and nineteenth-century fiction and drama and various databases of eighteenth- and nineteenth-century newspapers and periodicals.

It has been pointed out elsewhere (Beal 2012a) that the increasing availability of corpora compiled from texts of this period has revolutionised the study of Late Modern English in the twenty-first century. Denison notes that, "in the last two centuries, syntactic change has more often been statistical in nature, with a given construction... either becoming more or less common generally or in particular registers" (1998: 93). Since statistically-based studies require large amounts of comparable data, it is not surprising that Late Modern English scholarship has followed in the wake of Late Modern English corpora. The first decade of this century has seen the publication of three monographs dealing with the whole of this period (Beal 2004, C. Jones 2006, Tieken-Boon van Ostade 2009), as well as volumes dedicated to the eighteenth (Görlach 2001, Hickey 2010) and nineteenth centuries (Smitterberg 2005, Kytö et al. 2006). Furthermore, a series of conferences on Late Modern English, which began in Edinburgh in 2001, had its fifth meeting in Bergamo in 2013.

Whilst the above discussion seems to indicate that Late Modern English scholarship is in a healthy state, it has been argued (Beal 2012a) that phonology has been the poor relation in the Late Modern English family, largely due to the readier availability of corpora for the study of syntax and pragmatics. Although two monographs on Late Modern English pronunciation have been published (Beal 1999, C. Jones 2006), papers dealing with phonology have been in the minority in all the Late Modern English conferences held to date (see Beal 2012a: 22 for an analysis of the contents of publications from these conferences). The tendency for electronic corpora to be more useful for research in areas such as syntax and pragmatics is not confined to historical corpora. Anderson and Corbett point out that "most accessible online corpora focus on the printed word, even if occasionally these words have been annotated to show their pronunciation" (2009: 124). Nevertheless, several corpora of twentieth-century English pronunciation are now available, including the *Diachronic Corpus of Tyneside English* (DECTE), the *Phonologie d'Anglais Contemporain* (PAC) corpus, the *Scottish Corpus of Texts and Speech* (SCOTS), *A Sound Atlas of Irish English* (Hickey 2004) and the *Intonational Variation in English* (IViE) corpus, all of which allow the user to search sound files. Of course, sound files of eighteenth-

and most nineteenth-century speech are simply not available, so a corpus of historical English phonology would have to be based on printed information. In Section 2, we outline the nature of the evidence available for eighteenth-century English phonology and discuss its usefulness and suitability for corpus construction.

2. Evidence for eighteenth-century English phonology

Evidence for the pronunciation of English (or any other language) in historical periods preceding the invention of sound recording can be divided into two major categories: direct and indirect evidence (Beal 2012b: 63-64). Direct evidence consists of metalinguistic comments and linguistic descriptions from grammarians, lexicographers, orthöepists and others who are overtly and intentionally providing this information, whilst indirect evidence is pieced together from clues provided in rhymes, puns and spellings by authors who were almost certainly unaware that they were leaving phonological information for future historical linguists. Thus, as Beal points out:

> Shakespeare rhymed *war* with *jar* and *warm* with *harm* in *Venus and Adonis* (ll. 98/100 and 193/ 195 respectively) because he was writing within a tradition which demanded end-rhymes and because those words fitted in with the theme of his poem, not because he wished to record for posterity the fact that /w/ had not yet exerted a rounding influence on the following /a/. (1999: 37).

Shakespeare's rhymes thus provide indirect evidence for the unrounded pronunciation, and have been used as such by scholars such as Wyld (1923) and Kökeritz (1953). However, when the orthöepist Christopher Cooper (1687) provides a separate notation <α> for the vowel in *war, warden* and *warm* in a volume whose title page declares that it is "fitted for the Use of Schools and necessary for all those that desire to Read, Write or Speak our Tongue with Ease and Understanding" (1687: 1) he is deliberately providing this information for his contemporaries, and later generations of phonologists can deduce from this that the rounding had taken place by this date in the variety described by Cooper.

The balance of direct and indirect evidence for historical English pronunciation shifts from the Old and Middle English periods, from which direct evidence is very scarce, through the Early Modern (c. 1500-1700) period when, as we can see from the examples above, both kinds of evidence are plentiful, to the Late Modern period, when direct evidence predominates. Standardisation of spelling, increasing literacy and a greater acceptance of eye-rhymes in poetry meant that indirect evidence from this period became scarcer, whilst an increasing awareness of the social value of a 'correct' pronunciation created a market for pronouncing dictionaries and elocution manuals, especially in the second half of the eighteenth century. In his monograph on the eighteenth-century elocutionist

Thomas Sheridan, Benzie notes that “[F]ive times as many works on elocution were published between 1760 and 1800 than prior to 1760” (1972: 52). Dobson, whose major work on historical English pronunciation deals with the Early Modern period, dismissed eighteenth-century sources of direct evidence in the following sweeping statement:

> The eighteenth century produced no writers to compare either with the spelling reformers who are our main source up to 1644 (Hodges) or with the phoneticians who, beginning with Robinson (1617) carry us on from 1653 (Wallis) to 1687 (Cooper’s *English Teacher*). (Dobson 1957: 311)

However, as pointed out by Beal (1999: 47), Dobson was writing at a time when “the prevailing attitude was... that the study of English philology stopped at 1700” and the ease of access we now have to eighteenth-century texts via *ECCO* was unthinkable. Beal (1999) and C. Jones (2006) have since made extensive use of eighteenth-century sources to provide detailed accounts of the phonology of this period. Although the purpose of eighteenth-century elocutionists such as John Walker and Thomas Sheridan was undoubtedly prescriptive, Beal and C. Jones both demonstrate that their work can be taken seriously as providing evidence not only for what was considered the ‘correct’ pronunciation of their day, but also for pronunciations that were stigmatised and to be avoided. Furthermore, pronouncing dictionaries such as Sheridan (1780), Walker (1791) and many others from this period, provide descriptions of the recommended pronunciation of every word in the lexicon and thus, as Beal points out give us “invaluable detailed evidence of lexical diffusion” (1999: 68). As such, this evidence could be of use not only to historians of English, but to scholars in the fields of historical phonology more broadly and of language variation and change. However, as Beal (2007) has pointed out, there has been little use of eighteenth-century evidence by scholars researching the present-day diffusion of sound changes which began in that period. The provision of a searchable database of eighteenth-century phonology would greatly facilitate the use of the “past to explain the present” by researchers who may be unfamiliar with the complexities of eighteenth-century phraseology and notation. In the next sections, we will discuss the problems posed by these sources and propose a solution.

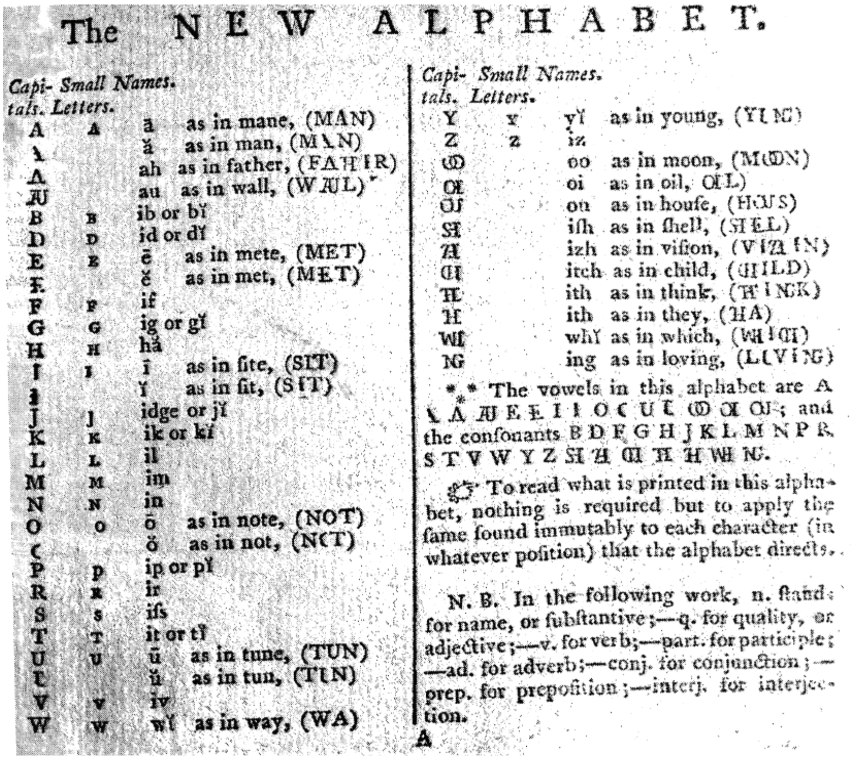
3. Problems arising from eighteenth-century sources

3.1 Problem 1: annotation

One major obstacle encountered by scholars embarking on research into Late Modern English phonology is the diversity of systems used by eighteenth- (and nineteenth-) century authors to represent the distinct sounds of English. The ubiquity of the International Phonetic Alphabet (IPA) in the twentieth century has made scholars reluctant to decipher earlier systems such as A. J. Ellis’s Palaeotype (see Local 1983, Maguire 2012 for discussion of Ellis’s system). Eighteenth-century authors, like the orthöepists of the sixteenth and seventeenth centuries, used a variety of methods to convey their recommended pronunciations to their readers. Abercrombie (1981) categorizes the orthographic systems used by these authors into two major schematic types: *new alphabets* and *augmentation of the Roman alphabet*, with the second type further subdivided into schemes using *diacritics* and *extended alphabets*. Although Thomas Spence described his system, illustrated in Figure 1, as a ‘New Alphabet’, according to Abercrombie’s scheme this would be categorized as an extended alphabet, based as it is on modification of the letters of the Roman alphabet. Even this was too radical for the majority of eighteenth-century readers, who preferred diacritic systems such as that exemplified by Walker’s ‘Table of the simple and diphthongal vowels’ (Figure 2). Here, the conventional spelling is not disrupted unless, as in words like *enough*, the pronunciation deviates considerably from that indicated by the usual values of the orthographic letters. In such cases, authors using diacritic systems would resort to semi-phonetic spelling: Walker represents this word as <e^{1}’-nu^{2}f>. This combination of semi-phonetic spelling and superscripted numbers to indicate separate vowel phonemes was first used in a pronouncing dictionary by Kenrick (1773), though the system had been described by Sheridan in his (1761) *Dissertation on the Causes of the Difficulties which Occur in Learning the English Tongue*. Sheridan went on to use this system in his (1780) *General Dictionary of the English Language*, and its adoption by Walker ensured that this would be the most successful and widespread system of the eighteenth and early nineteenth centuries. However, each author who uses this system has his own way of representing specific sounds, so that an <a> with superscripted <1> has a different phonetic value in Walker’s dictionary, where it represents /eː/ as in *fate*, and in Sheridan’s, where it represents /a/ as in *hat*.

It should be apparent from the above discussion that the existence of such a variety of notation systems would prove an obstacle to the comparison of pronunciations recommended by different authors of the Late Modern period. The researcher must decipher each system and translate each combination of symbol and diacritic into IPA in order to make such comparisons. Those who have undertaken such projects (Beal 1999, C. Jones 2006, MacMahon 1998) have had to search each source manually to make these comparisons.

Beal (1999) created a searchable database of all the entries in Thomas Spence’s *Grand Repository of the English Language* (1775) by recoding them from Spence’s ‘New Alphabet’ into alphanumeric characters, as set out in Figure 3. She then used the Oxford Concordance Programme (OCP) to generate lists of words containing specific symbols in specified environments, which, given the phonemic nature of Spence’s system, provided all instances of a particular phoneme environment in Spence’s lexicon. Each word in the list was then looked up in a number of other eighteenth-century pronouncing dictionaries, including Sheridan (1780) and Walker (1791) to yield evidence of variation and change,



The NEW ALPHABET.

Capi-tals.	Small Letters.	Names.	
A	A	ā	as in mane, (MAN)
Ʌ		ă	as in man, (MɅN)
Δ		ah	as in father, (FΔÐƗR)
Æ		au	as in wall, (WÆL)*
B	B	ib or bĭ	
D	D	id or dĭ	
E	E	ē	as in mete, (MET)
Ɇ		ĕ	as in met, (MɆT)
F	F	if	
G	G	ig or gĭ	
H	H	hă	
I	I	ī	as in ſite, (SIT)
Ɨ		ĭ	as in ſit, (SƗT)
J	J	idge or jĭ	
K	K	ik or kĭ	
L	L	il	
M	M	im	
N	N	in	
O	o	ō	as in note, (NOT)
Ɔ		ŏ	as in not, (NƆT)
P	p	ip or pĭ	
R	R	ir	
S	S	iſs	
T	T	it or tĭ	
U	U	ū	as in tune, (TUN)
Ʊ		ŭ	as in tun, (TƱN)
V	V	iv	
W	W	wĭ	as in way, (WA)
Y	Y	yĭ	as in young, (YƱŊ)
Z	Z	iz	
Ꝏ		oo	as in moon, (MꝎN)
Œ		oi	as in oil, ŒL)
Ȣ		ou	as in houſe, (HȢS)
Ʃ		iſh	as in ſhell, (ƩɆL)
Ʒ		izh	as in viſion, (VƗƷƗN)
Ꞓ		itch	as in child, (ꞒILD)
Þ		ith	as in think, (ÞƗŊK)
Ð		ith	as in they, (ÐA)
Ƕ		whĭ	as in which, (ǶƗꞒ)
Ŋ		ing	as in loving, (LƱVƗŊ)

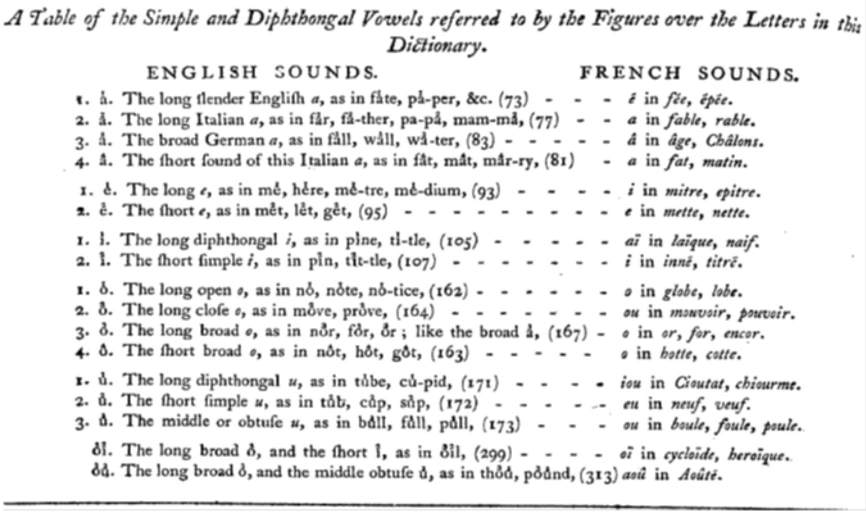
*** The vowels in this alphabet are A Ʌ Δ Æ E Ɇ I Ɨ O Ɔ U Ʊ Ꝏ Œ Ȣ; and the conſonants B D F G H J K L M N P R S T V W Y Z Ʃ Ʒ Ꞓ Þ Ð Ƕ Ŋ.

☞ To read what is printed in this alphabet, nothing is required but to apply the ſame ſound immutably to each character (in whatever poſition) that the alphabet directs.

N. B. In the following work, n. ſtands for name, or ſubſtantive;—q. for quality, or adjective;—v. for verb;—part. for participle;—ad. for adverb;—conj. for conjunction;—prep. for prepoſition;—interj. for interjection.

A

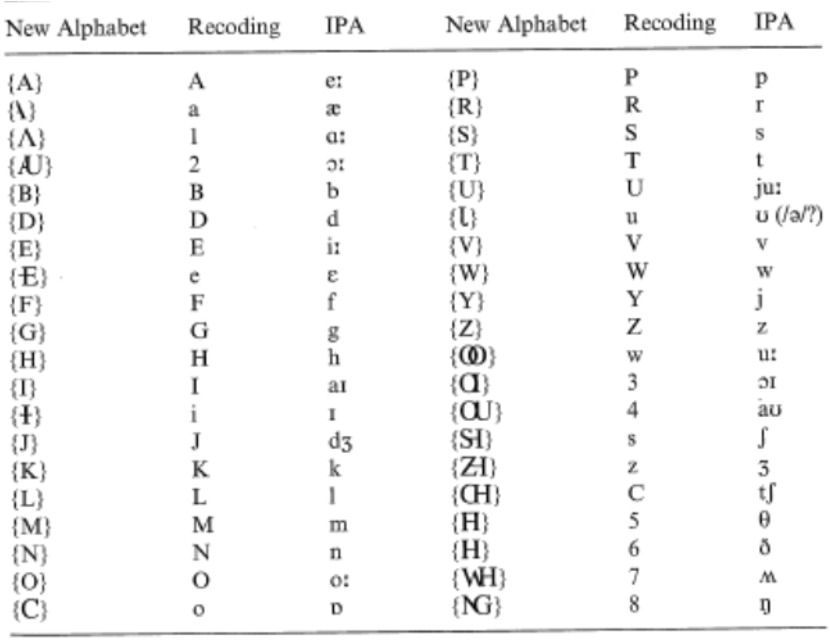
Figure 1. Spence's 'New Alphabet'



A Table of the Simple and Diphthongal Vowels referred to by the Figures over the Letters in this Dictionary.

		ENGLISH SOUNDS.	FRENCH SOUNDS.
1.	$\overset{1}{a}$.	The long ſlender Engliſh *a*, as in f$\overset{1}{a}$te, p$\overset{1}{a}$-per, &c. (73)	*é* in *fée, épée.*
2.	$\overset{2}{a}$.	The long Italian *a*, as in f$\overset{2}{a}$r, f$\overset{2}{a}$-ther, pa-p$\overset{2}{a}$, mam-m$\overset{2}{a}$, (77)	*a* in *fable, rable.*
3.	$\overset{3}{a}$.	The broad German *a*, as in f$\overset{3}{a}$ll, w$\overset{3}{a}$ll, w$\overset{3}{a}$-ter, (83)	*â* in *âge, Châlons.*
4.	$\overset{4}{a}$.	The ſhort ſound of this Italian *a*, as in f$\overset{4}{a}$t, m$\overset{4}{a}$t, m$\overset{4}{a}$r-ry, (81)	*a* in *fat, matin.*
1.	$\overset{1}{e}$.	The long *e*, as in m$\overset{1}{e}$, h$\overset{1}{e}$re, m$\overset{1}{e}$-tre, m$\overset{1}{e}$-dium, (93)	*i* in *mitre, epitre.*
2.	$\overset{2}{e}$.	The ſhort *e*, as in m$\overset{2}{e}$t, l$\overset{2}{e}$t, g$\overset{2}{e}$t, (95)	*e* in *mette, nette.*
1.	$\overset{1}{i}$.	The long diphthongal *i*, as in p$\overset{1}{i}$ne, t$\overset{1}{i}$-tle, (105)	*aï* in *laïque, naif.*
2.	$\overset{2}{i}$.	The ſhort ſimple *i*, as in p$\overset{2}{i}$n, t$\overset{2}{i}$t-tle, (107)	*i* in *inné, titré.*
1.	$\overset{1}{o}$.	The long open *o*, as in n$\overset{1}{o}$, n$\overset{1}{o}$te, n$\overset{1}{o}$-tice, (162)	*o* in *globe, lobe.*
2.	$\overset{2}{o}$.	The long cloſe *o*, as in m$\overset{2}{o}$ve, pr$\overset{2}{o}$ve, (164)	*ou* in *mouvoir, pouvoir.*
3.	$\overset{3}{o}$.	The long broad *o*, as in n$\overset{3}{o}$r, f$\overset{3}{o}$r, $\overset{3}{o}$r; like the broad $\overset{3}{a}$, (167)	*o* in *or, for, encor.*
4.	$\overset{4}{o}$.	The ſhort broad *o*, as in n$\overset{4}{o}$t, h$\overset{4}{o}$t, g$\overset{4}{o}$t, (163)	*o* in *hotte, cotte.*
1.	$\overset{1}{u}$.	The long diphthongal *u*, as in t$\overset{1}{u}$be, c$\overset{1}{u}$-pid, (171)	*iou* in *Cioutat, chiourme.*
2.	$\overset{2}{u}$.	The ſhort ſimple *u*, as in t$\overset{2}{u}$b, c$\overset{2}{u}$p, ſ$\overset{2}{u}$p, (172)	*eu* in *neuf, veuf.*
3.	$\overset{3}{u}$.	The middle or obtuſe *u*, as in b$\overset{3}{u}$ll, f$\overset{3}{u}$ll, p$\overset{3}{u}$ll, (173)	*ou* in *boule, foule, poule.*
	$\overset{3}{o}\overset{2}{i}$.	The long broad $\overset{3}{o}$, and the ſhort $\overset{2}{i}$, as in $\overset{3}{o}\overset{2}{i}$l, (299)	*oï* in *cycloïde, heroïque.*
	$\overset{3}{o}\overset{3}{u}$.	The long broad $\overset{3}{o}$, and the middle obtuſe $\overset{3}{u}$, as in th$\overset{3}{o}\overset{3}{u}$, p$\overset{3}{o}\overset{3}{u}$nd, (313)	*aoû* in *Aoûté.*

Figure 2. Walker's 'Table of Simple and Diphthongal Vowels'



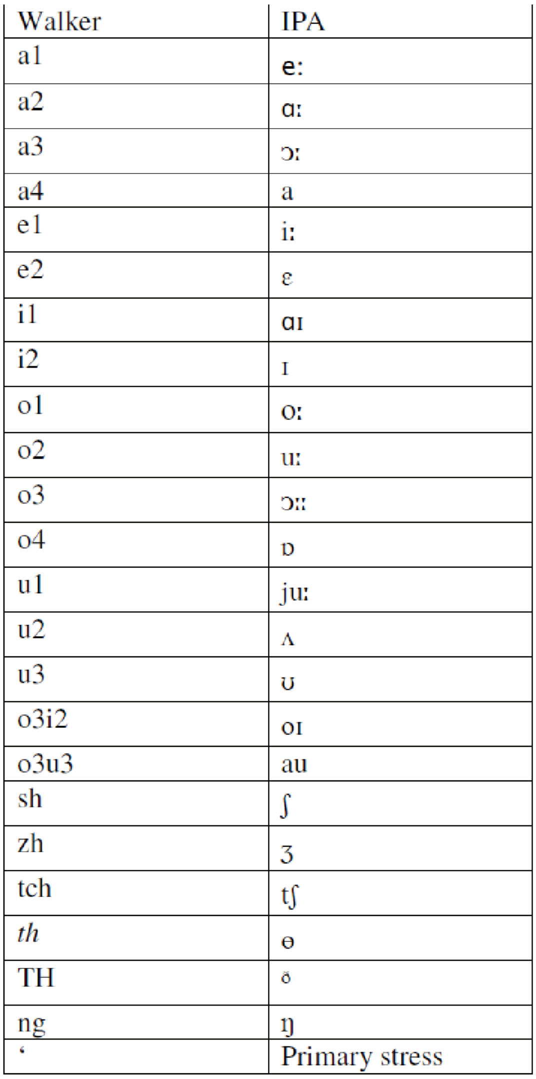
New Alphabet	Recoding	IPA	New Alphabet	Recoding	IPA
{A}	A	eː	{P}	P	p
{\}	a	æ	{R}	R	r
{Λ}	1	ɑː	{S}	S	s
{ɅU}	2	ɔː	{T}	T	t
{B}	B	b	{U}	U	juː
{D}	D	d	{Ɩ}	u	ʊ (/ə/?)
{E}	E	iː	{V}	V	v
{Ɇ}	e	ɛ	{W}	W	w
{F}	F	f	{Y}	Y	j
{G}	G	g	{Z}	Z	z
{H}	H	h	{ꝏ}	w	uː
{I}	I	aɪ	{CI}	3	ɔɪ
{Ɨ}	i	ɪ	{CU}	4	aʊ
{J}	J	dʒ	{SH}	s	ʃ
{K}	K	k	{ZH}	z	ʒ
{L}	L	l	{CH}	C	tʃ
{M}	M	m	{H}	5	θ
{N}	N	n	{H}	6	ð
{O}	O	oː	{WH}	7	ʍ
{C}	o	ɒ	{NG}	8	ŋ

Figure 3. Recording of Spence's 'New Alphabet'

including lexical diffusion. Beal notes that this task was "painstaking and time consuming", and, whilst it yielded a great deal of useful information, the study "barely scratched the surface in terms of the wealth of phonological evidence available in eighteenth-century pronouncing dictionaries" (1999: 183-184).

The system set out in Figure 3 was devised by Beal on an *ad hoc* basis at a time when it would have been difficult to find an IPA font compatible with OCP. For the proposed database, the combinations of character and diacritic denoting specific phonemes in each eighteenth-century source will be transliterated into the Unicode equivalent. Figure 4 shows the notations used by Walker (1791), with their equivalents in Unicode IPA.

It is important to note that the notational equivalents in Figure 4 are intended to be phonemic: the transliteration of Walker's <a^2> as IPA /ɑː/ is not intended to suggest that the vowel concerned had the same (back) articulation in Walker's time as in present-day RP, simply that it is a separate phoneme from Walker's <a^4>. The creation of a database of eighteenth-century phonology will inevitably require us to make decisions concerning the attribution of notations in the historical sources to their equivalent Unicode phonemic notations, but all such decisions will be accounted for in the metadata accompanying the database. In Section 3.2, we will discuss the size of this proposed database.



Walker	IPA
a1	eː
a2	ɑː
a3	ɔː
a4	a
e1	iː
e2	ɛ
i1	ɑɪ
i2	ɪ
o1	oː
o2	uː
o3	ɔːː
o4	ɒ
u1	juː
u2	ʌ
u3	ʊ
o3i2	oɪ
o3u3	au
sh	ʃ
zh	ʒ
tch	tʃ
th	θ
TH	ð
ng	ŋ
‘	Primary stress

Figure 4. Walker’s (1791) Notation Transliterated into IPA

3.2 Problem 2: size and scope

Beal (1999) was able to transliterate the whole of Spence’s (1775) dictionary because it is relatively short, consisting of approximately 17,000 entries. A pilot study carried out in 2010 established that a highly competent research assistant was able to transliterate 3,378 entries from Walker (1791) over a period of 40 hours.[1] This only covered the entries from *abacus* to *borage*, indicating that it would take a great deal of time, and therefore expense, to include every word from every eighteenth-century source in the proposed database. Whilst, as Beal (1999) has demonstrated, access to a complete lexicon does provide valuable evidence of lexical diffusion, many of the words recorded in eighteenth-century dictionaries are obscure and/or now obsolete. Examples from the pilot project include *arundinacious, atrabilariousness, belswagger,* and *bezoardick*, all magnificent words but unlikely to be amongst those included in studies of English historical phonology (unless, of course, the object of the research was an investigation of stress patterns in polysyllabic words). In order to keep the database to a manageable size, we propose to restrict the entries to the words used by Wells (1982) to illustrate his standard lexical sets. In Wells's system each keyword "stands for a large number of words which behave the same way in respect of the incidence of vowels in different accents" (1982: 120). Since this system also includes subsets which differentiate between historical lexical sets, it is as useful for diachronic as for diatopic comparisons.

Including all the words provided by Wells in subsets of lexical sets would give 1,739 items, about one tenth of the size of Spence's (1775) dictionary, but we can see from the example of the FLEECE set in Figure 6 that not all of these would be included in eighteenth-century pronouncing dictionaries. Although some of these dictionaries did include proper names, *Keith,* and *Sheila* are unlikely to appear (though *Peter* may be in some); likewise *casino,* and *ski* are first cited in 1789 and 1755 respectively in the *Oxford English Dictionary*. On the other hand, since Wells's lexical sets are designed for the comparison of vowel phonemes and their distribution, further sets will need to be provided if users of the database are to have access to information concerning consonantal variants, such as /hw ~w/ discussed below. Subtracting from Wells's list such words as do not appear in the eighteenth-century sources and augmenting it with a small number of consonantal sets would yield a database of manageable size which would nevertheless provide a rich amount of information on the diachronic, diatopic and lexical distribution of phonological variants in eighteenth-century English. The next section consists of a case study in which a supplementary lexical set for /hw ~w/ was compared across a subset of nine eighteenth-century pronouncing dictionaries to reveal patterns of variation and change.

4. A pilot study: the eighteenth-century pronunciation of 'wh'

4.1 Method

We described above plans for a searchable database of eighteenth-century English phonology. In this section, we test whether such a resource might usefully answer questions about phonological variation and change.

	1757	1772	1773	1775	1775	1780	1786	1791	1797	1701- 1800
	BUCHANAN	**JOHNSTON**	**KENRICK**	**SPENCE**	**PERRY**	**SHERIDAN**	**BURN**	**WALKER**	**JONES**	**OED COUNT**
WHALE	hw	hw	w	hw	w	hw	w	hw	hw	84
WHARF	hw	w	w	w	w	hw	w	hw	hw	15
WHAT	hw	hw	w	hw	w	hw	w	hw	hw	2611
WHEAT	hw	hw	w	hw	w	hw	w	hw	hw	157
WHEEDLE	hw	hw	w	hw	w	hw	w	hw	hw	2
WHEEL	hw	hw	w	hw	w	hw	w	hw	hw	190
WHEEZE	hw	hw	w	hw	w	hw	w	hw	hw	5
WHELM	hw	hw	hw	hw	hw	hw	w	hw	hw	0
WHELP	hw	hw	w	hw	w	hw	w	hw	hw	5
WHEN	NA	hw	w	hw	w	hw	w	hw	hw	3742
WHENCE	NA	hw	w	hw	w	hw	w	hw	hw	264
WHERE	NA	hw	w	hw	w	hw	w	hw	hw	1900
WHERRY	hw	hw	hw	hw	w	hw	w	hw	hw	6
WHET	hw	hw	w	hw	w	hw	w	hw	hw	5
WHETHER	hw	hw	hw	hw	w	hw	w	hw	hw	629
WHEY	hw	hw	w	hw	w	hw	w	hw	hw	15
WHICH	NA	hw	w	hw	w	hw	w	hw	hw	9201
WHIFF	hw	hw	w	hw	w	hw	w	hw	hw	8
WHIFFLE	hw	hw	w	hw	w	hw	w	hw	hw	0
WHIG	hw	hw	w	hw	w	hw	w	hw	hw	69
WHILE	NA	hw	w	hw	w	hw	w	hw	hw	852
WHIM	hw	hw	w	hw	w	hw	w	hw	hw	18
WHIMPER	hw	hw	w	hw	w	hw	w	hw	hw	2
WHIN	hw	NA	w	hw	w	hw	w	hw	hw	13
WHINE	hw	hw	w	hw	w	hw	w	hw	hw	11
WHIP	hw	hw	w	hw	w	hw	w	hw	hw	57
WHIRL	hw	hw	w	hw	w	hw	w	hw	hw	29
WHISK	hw	hw	hw	hw	hw	hw	w	hw	hw	14
WHISKERS	hw	hw	hw	hw	hw	hw	w	hw	hw	20
WHISPER	hw	hw	hw	hw	hw	hw	w	hw	hw	29
WHIST	hw	hw	hw/ w	hw	hw/ w	hw	w	hw	hw	31
WHISTLE	hw	hw	hw	hw	w	hw	w	hw	hw	60
WHIT	hw	hw	w	hw	w	hw	w	hw	hw	16
WHITE	hw	hw	w	hw	w	hw	w	hw	hw	1621
WHITHER	hw	hw	w	hw	w	hw	w	hw	hw	12
WHITLOW	hw	hw	hw	hw	hw	hw	hw	hw	hw	1

WHITSUNTIDE	hw	hw	hw	hw	w	hw	hw	hw	hw	2
WHIZ	hw	hw	hw	hw	hw	hw	hw	hw	hw	4
WHO	hw	h	h	hw	h	h	h	h	h	4333
WHOLE	h	h	h	hw	h	h	h	h	h	1221
WHOM	NA	NA	h	hw	h	h	h	h	h	475
WHOOP	hw/w	h	h	hw	h	h	h	h	h	11
WHORE	h	h	h	h	h	h	h	h	h	9
WHOSE	NA	h	h	hw	h	h	h	h	h	986
WHY	NA	hw	w	hw	w	hw	w	hw	hw	392
SOMEWHERE	NA	hw	NA	hw	w	hw		hw	hw	49
SOMEWHAT	NA	NA	NA	hw	w	hw		hw	hw	484
OVERWHELM	hw	hw	NA	hw	hw	hw		hw	hw	5
NOWHERE	NA	NA	NA	hw	w	hw		hw	hw	39
ELSEWHERE	NA	NA	NA	hw	hw	hw		hw	hw	140

Figure 5. Data from pilot study

Our test case involves the representation of ‘wh’ in nine eighteenth-century pronouncing dictionaries. In present-day standard southern British English, words such as *whale*, *what*, and *where* begin with /w/, whilst *who* and *whole* have initial /h/. Eighteenth-century sources present evidence, through their orthographic systems, of variation across authors between /hw/ and /w/ for the first set, hence a preserved versus unpreserved /hw ~ w/ contrast in *where ~ wear*. The nine dictionaries were selected to ensure that variation in pronunciation as well as in geography and chronology were amply represented. We recorded in a spreadsheet the pronunciations of 50 words which occur in as many as possible of the nine dictionaries, consisting of (1) 39 words beginning with the spelling ‘wh’ which are pronounced with /w/ in present-day southern British English, (2) 6 words with initial ‘wh’ which are now pronounced with initial /h/, and (3) 5 words with ‘wh’ word internally, which are now all pronounced with internal /w/. The nine authors were arranged as columns in chronological order, with an additional column displaying the total of how many times each word appears in quotations used by the OED dating from the eighteenth century (1701-1800), to give an indication of their frequency. The words under consideration were listed as rows. Figure 5 presents the evidence as described.

This systematic data collection even on such a small scale enabled us to identify patterns in the evidence, along dimensions commonly under investigation in sociolinguistic, historical and phonological research, namely geography, chronology, phonology, lexical factors, and social class. Furthermore, the nature of the data also enabled us to glean ‘direct’ evidence in the form of contemporary commentary on the choices made by the authors. A notable example is that Walker presents the loss of the /hw ~ w/ contrast as a special case of ‘h-dropping’, which was just beginning to attract social stigma at this time in lower-class London English. The proposed database would include such information.

As our study aims to ascertain and explain the variation in pronunciation of ‘wh’ in the eighteenth century, it is first useful to present background research on two aspects: firstly, a reconstruction of the nature of the /hw ~ w/ contrast going into the eighteenth century, and secondly, the phonetic nature of the sound or cluster we are treating as /hw/.

4.2 The starting point: /hw/ before the eighteenth century

Words containing /hw/ in English (< Old English hw < Common Germanic *xw < Proto-Indo-European *kw) had already begun to be pronounced with simple /w/ in the twelfth century in many southern dialects, notably in London (Dobson 1957: 974). However, /hw/ was clearly not unknown in southern speech for many centuries, as shown by the fact that spellings with simple <w> are much sparser than would be expected in the fifteenth to seventeenth centuries if this was the regular pronunciation (Wyld 1936: 312). Johnston (1764: 9) comments that the ‘h’ element in these words was at the time “very little heard”, which appears to indicate, from the context, that these forms had weak aspiration in normal speech, and not that *few people* pronounced them as /hw/. Contrary to the southern position, most northern English and Scottish dialects robustly preserved /hw/, which persists to this day in Scottish dialects.

The development of */hw/ to /h/ before back, rounded vowels such as /u/ (e.g. *who*) seems to date from the thirteenth or fourteenth centuries, but only entered conservative registers in the seventeenth century (Dobson 1957: 980-981). The /h/ pronunciation was reasonably settled in southern England by the eighteenth century, but data from north-east England (Spence 1775) suggests that /hw/ persisted in these dialects for longer (see Section 4.44).

Therefore, entering the eighteenth century, the /hw ~w/ contrast was only weakly realised in southern English, and the /hw ~ h/ contrast before back, rounded vowels no longer realised, whereas /hw ~ w/ was robustly preserved in northern English and Scottish dialects, and there is evidence to indicate that /hw ~ h/ also remained in some northern English varieties.

4.3 The phonetics of /hw/

The phonological nature of ‘wh’ (when not simply /w/) in eighteenth-century English is unclear: it could either be analysed as a consonant cluster /hw/ or a single voiceless labial-velar fricative or approximant /ʍ/. The question is discussed further in Section 4.4 and Section 4.7, but generally the notation /hw/ is employed throughout this pilot study. In either case, there are labial and velar place-of-articulation elements, and a breathy, aspirated element which may or may not produce audible frication, hence the phonetic realisation of the phonologically different representations need not be different. Ladefoged and Maddieson (1996: 326) report that, in the world’s languages, /ʍ/ is usually non-fricative, but go on to acknowledge that a fricative realisation is a possibility, in which case it is "better described as a voiceless labialized velar fricative", namely [x^w], as "the voiceless counterpart of w cannot have friction at both the labial and velar places of articulation". Reconstructing such a phonetic realisation of the phoneme or cluster would be consistent with one historical observation and one piece of contemporary commentary. Historically, greater restrictions seem to have developed between /hw/ and the labiality of the *following* vowel rather than the preceding, an asymmetry which would be predicted by a labializing secondary articulation, as Ladefoged and Maddieson (1996: 357) note that a labial articulatory gesture is anchored in terms of its timing to the release rather than the formation of a primary articulation. Hence, we have */hw/ > /h/ in words such as *who*, where a vowel with lip-rounding follows. Secondly, Douglas ([1779] 1991: 141) comments that "The Scottish pronounce the *wh* like their guttural *ch*", and "When they endeavour to correct this fault they are apt to omit the *h*, so as to pronounce *whit* and *wit*... in the very same manner". Given that the "fault" appears to be the production of velar frication, it would be predicted that the omission of such an element would result in precisely a labial-velar approximant, thus /hw/*it* > /w/*it*. However, Douglas' comment does suggest that the difference between *whit* and *wit* ought then to be preserved in a different way from the production of strong velar frication, suggesting two alternatives: weak velar frication or labial frication.

A logical alternative realisation of /hw/ or /ʍ/ would be a voiceless velarized labial fricative [ɸ^ɣ], i.e. frication produced at the lips, not the velum, and this is the pronunciation concluded by M. Jones (2008) for the present-day Scottish variant, based on acoustic and articulatory evidence. Jones identifies weak frication produced at the lips, as suggested by the generally flat acoustic spectrum, and the fact that lip aperture is different between /w/ and /ʍ/, indicating a different modification of the airflow at the lips between the two sounds.[2] The acoustic formats are generally those for a labial-velar, indicating that there is an approximant-degree constriction at the velum in addition to the labial frication. Labial frication historically in Scotland appears to be corroborated by the development of */hw/ to /f/ in north-east Scottish dialects, a change that is only plausible if we posit that the velar element of the sound was lost due to its minimal acoustic effect.

We should therefore bear in mind when analysing the eighteenth-century data that there may be geographical variation in the phonetic realisation of the sound represented by 'wh'.

4.4 Analysing the data 1: geography

Three main patterns emerge from the data based on geographical distribution: (1) The London authors prefer /hw/ to /w/, with the exception of Kenrick, (2) Two out of the three Scottish authors prefer /w/ (Perry 1775, Burn 1786), whereas the earliest, Buchanan (1775) prefers /hw/, and (3) Spence, from Newcastle, has /hw/ even in words containing a following back, rounded vowel, e.g. *who*; all the other authors have /h/ in this position with exceptions discussed in Section 4.7 below.

Sheridan, from Ireland, has /hw/ consistently, and /h/ where expected, but he patterns in many ways with the London-based authors, as discussed below.

The London authors' preference for /hw/ appears to be inconsistent with the conclusion in Section 4.2 above that /hw/ was only weakly realised in southern English at the start of the eighteenth century. We might have expected the /hw ~w/ contrast to be further eroded and not identified by the dictionary authors in the second half of the century, and indeed most indirect sources of evidence suggest that /w/ was the norm in London at this time. However, the near-consistent transcription using /hw/ appears to reflect a prescriptive attempt to revitalise this contrast under the influence of spelling, on the basis that the simplification of /hw/ to /w/ could be considered a form of 'h-dropping', that is, the common omission of the initial fricative in words such as *happy* in the lower-class London English pronunciation of the time (Beal 1999: 176-178). This phenomenon was beginning to attract social stigma in the middle of the eighteenth century, and was one precisely proscribed by London-based authors such as Walker and S. Jones, who both explicitly classed the /w/ pronunciation as 'h-dropping', and made overt comments labelling this practice a vulgarism. The Irishman Sheridan, who spent a number of years in London, also specifically proscribed the /w/ pronunciation, aligning it with the stigmatized 'h-dropping'. This account of the London pattern is supported by the fact that the only London-based exception to /hw/ is Kenrick (1773), one of the earliest of the group; presumably the stigmatization of /w/ had not yet fully taken effect by this time. The otherwise consistent /hw/ pronunciation in the dictionaries instead of the apparently regular London /w/ can therefore be considered 'collateral damage' from the stigma of 'h-dropping'.

The preference for /w/ among the Scottish authors is again curious in the light of the conclusions in Section 4.22 and Section 4.3 above that /hw/ remained the regular Scottish pronunciation throughout this period and indeed to the present day. Recall, however, Douglas' (([1779] 1991: 141)) comments that "The Scottish pronounce the *wh* like their guttural *c"*", and "When they endeavour to correct this fault they are apt to omit the *h*". This observation is consistent with Perry's and Burn's near-consistent /w/ (the exceptions are discussed below in Section 4.6.2, Section 4.7.3, and Section 4.7.4), if we posit that these authors were advising a more London-like pronunciation to avoid the Scottish /hw/, stigmatized due to its clear regional connotations. The /w/ pronunciation could therefore be analysed as a hypercorrect Anglicism, one which is particularly remarkable in the light of the contemporaneous opposite trend in London where /hw/ was proscribed due to 'h-dropping'. Arguably, this trend was only taking hold in London at the time and had not yet reached the consciousness of the Scottish authors. Finally, and similarly to the London pattern, the exception to the Scottish pattern is the earliest from that geographical area, Buchanan (1757), who has near-consistent /hw/, presumably reflecting the standard Scottish pronunciation of the time and up to the present day.

Spence (1775) from Newcastle in north-east England has extremely consistent /hw/, even in words containing a following back, rounded vowel (the only exceptions are *wharf* and *whore*, discussed below in Section 4.7). One can reasonably assume that Spence's transcription accurately reflects the regional pronunciation of the time (Beal 1999: 179-180). Furthermore, Spence is the only author to use a special symbol for the /hw/ sound: an upper-case WH ligature. This could indicate that in Spence's dialect, the sound was not a consonant cluster /hw/, but a single phoneme /ʍ/, and this account is further supported by the preservation of the labial element even in words such as *who*, *whole*, *whom*, *whoop* and *whose*, all of which Spence transcribes using his ligature. The preservation of the labial element in a single phoneme would then be entirely in parallel with its preservation in the *voiced* labial-velar phoneme /w/ before a back, rounded vowel in almost all English dialects, including those of north-east England, as in *wound*, *womb*, *wool*, *wood*, etc. Undoubtedly a single phoneme, /w/ did not delete or develop to a velar or glottal sound in this environment. Also analysing /ʍ/ as a single phoneme in Spence's Newcastle dialect explains this identical behaviour, and implies that in most other English dialects – which have /h/ before back, rounded vowels – the correct phonological analysis of the 'wh' sound is a cluster /hw/. Delabialisation could then be analysed as cluster simplification, with loss of the /w/ element before a back, rounded vowel, in parallel with the widespread loss of /w/ in other clusters in this exact environment, e.g. *sword*, where */sw/ > /s/ several centuries earlier.[3] Recall from Section 4.2 that */hw/ > /w/ in this environment began in the thirteenth to fourteenth centuries.

To summarise, the collection of evidence from several eighteenth-century dictionaries, both from pronunciations and commentaries, allows us to reconstruct and explain the geographical distribution of the variant realisation of the 'wh' sound, from London to Newcastle in England, through to Scotland.

4.5 Analysing the data 2: chronology

The chronological patterns for the distribution of 'wh' have been identified above: Kenrick, a London author from an earlier period of the time-frame, has a /w/ pronunciation because the stigma of 'h-dropping' and the identification of /w/ with this phenomenon was only recently taking hold in educated London speech. Buchanan, the earliest Scottish author in the pilot study, has a /hw/ pronunciation, unlike the later Scottish authors, who are presumably recommending a more London-like, and less 'guttural', pronunciation, unaware that trends in London were simultaneously changing. In future studies using a more complete database, the chronological dimension of the problem under investigation will be easily visualised and notable patterns more readily extracted.

4.6 Analysing the data 3: lexical factors

4.6.1 Homophones

Certain authors make use of the /hw ~ w/ contrast to differentiate the two meanings of the words *whist* and *whoop*, constructing minimal pairs from the contrast. Kenrick and Perry both have /hw/ for *whist* 'be quiet', but /w/ for *whist* 'card game'. It is perhaps not coincidental that these are the two authors (in the pilot study) who are most variable in their selection of /hw/ or /w/: both generally prefer /w/, but a number of unexpected /hw/ words appear, discussed further in Section 4.6.2 and Section 4.7.3 below. Both were arguably sensitive to the contrast, and presumably believed there to be differences between lexical items as to which was the correct sound (i.e. not an 'across-the-board' /hw/ or /w/). Similarly, Buchanan has /hw/ for *whoop* 'a cry', but /w/ for *whoop* 'a bird'. Buchanan elsewhere consistently has /hw/, suggesting that the contrast is being used solely to differentiate between words that would otherwise be homophones.

A near-minimal pair can also be identified in Burn's and Perry's pronunciations for *Whitsuntide* with /hw/ and *whit* with /w/. Note that both authors were Scottish, but both generally preferred /w/, presumably in conflict with the regular pronunciations of their regions. The construction of a minimal pair therefore suggests sensitivity to the contrast.

4.6.2 Onomatopoeia and sound symbolism

The /hw/ pronunciation for *whist* 'be quiet' and *whoop* 'a cry' reported above could also be analysed as examples of onomatopoeic pronunciation. Further plausible examples of onomatopoeia or sound symbolism again come from Kenrick and Perry:

(1) Kenrick's /hw/ pronunciations analysable as onomatopoeia/symbolism
whisk, whisper, whistle, whiz

(2) Perry's /hw/ pronunciations analysable as onomatopoeia/symbolism
whisk, *whisper*

It is perhaps also not a coincidence that these words all have a relatively high, front vowel following the /hw/ sound, as such a tongue position might plausibly enhance velar frication (see Section 4.7.3).

4.6.3 Word frequency

In this pilot study, we attempted to gauge the frequency of the words under consideration to ascertain whether any pattern emerged. Frequency has been shown to play an important role in the spread of sound changes in the lexical diffusion model (e.g. Philips 1984). If our data enables us to catch a change midstream, whether that change is /hw/ > /w/ by regular sound change, or /w/ > /hw/ due to prescriptive or social factors, word frequency might show us how far the change has progressed, if the change is lexically diffused. If the change is not of this type, we might expect frequency to play a minimal role. Incorporating word frequency information in any database would therefore provide a valuable tool to test theories of sound change.

We compiled a word count of all the forms under consideration in quotations used by the *Oxford English Dictionary* during the time-period 1701-1800, to give a rough indication of the frequencies of the words at the time. Few patterns emerge from the word count, other than that Kenrick's and Perry's unexpected /hw/ words which are not onomatopoeic have very low counts: *whelm* has zero and *wherry* only six. Compare these figures with words containing /w/ in these dictionaries: *whale* has a count of 84, and *whig* 69. However, the pattern is not wholly corroborated by the data, in that other apparently low-frequency items have the /w/ pronunciation, such as *wheedle* (two) and *wheeze* (five). It is possible that a more complete word count will provide clearer patterns or a clearer indication of their absence, to enable us to ascertain whether frequency played a role. We therefore intend to use ARCHER to extend the word-count information for the final database.

4.7 Analysing the data 4: phonology

4.7.1 Before high, rounded vowels

The clearest phonological environment conditioning the distribution of 'wh' variants is of course before a high, rounded vowel. From the forms in the data, it appears that any vowel that is higher and more round than /ɔ/ on the vowel quadrilateral results in a realisation of the sounds written 'wh' as /h/, and not /hw ~ w/, and this is presumably due to a labial co-occurrence constraint. In Section 4.4, we concluded that this might best be analysed as deletion of /w/ in a /hw/ cluster, as dialects where the sound was arguably a phonological singleton (Spence's Newcastle dialect) resisted the deletion. The conclusion that the vowel had to be higher than /ɔ/ arises from the absence of /h/ realisations in *wharf* in any of the dictionaries.

Whore consistently has a pronunciation with /h/ in every one of the nine dictionaries in the pilot study, despite (1) possibly having a vowel /ɔ/, although OED reports that the eighteenth-century pronunciation was (the now dialectal) /hʊə/, and (2) Spence having no /h/ realisations, except in this instance. The solution appears to be historical: *whore* etymologically has initial /h/ (Old English *hore*), not /hw/, hence the pronunciation variant with /hw/ arguably never arose.

A second instance where etymology provides a solution is Buchanan's choice of /hw/ for *who* (like Spence), but /h/ for *whole* (unlike Spence). Not only does *whole* etymologically have /h/ (Old English *hāl*), furthermore the standard spellings of *whole* in Scots were *hail* and *hale*. Hence presumably the Scottish

Buchanan considered the /h/ pronunciation the robust standard. Buchanan's other interesting choice in this phonological environment is in the *whoop* minimal pair with /hw/ and /w/ discussed in Section 4.6 above.

4.7.2 Wharf

The initial sound in *wharf* follows whichever practice is generally adopted by seven of the nine authors (e.g. /hw/ in Walker, /w/ in Burn), but both Spence and Johnston rather curiously have /w/ instead of the usual /hw/ given by these two authors. The occurrence here of /w/ in Spence is particularly striking, given that there are no other such pronunciations of 'wh' in his dictionary, and he has /hw/ even in words with a following high, rounded vowel (*whore* being the only exception, as above). How can this peculiarity be explained?

Clearly, the rounding of the vowel is not a relevant factor, as a labial dissimilation would presumably have yielded /h/, not /w/, and it is clear that such phenomenon does not take place in Spence's dialect Section 4.4). We must discover an explanation that accounts for the loss of the *fricative* element. We identified in Section 4.3 that the alternatives for the phonetic realisation of /hw/ or fricated /ʍ/ were (1) a labialized velar fricative [x^{w}], or (2) a velarized labial fricative [ɸ^{ɣ}]. In account (1), frication is produced by a high tongue back, so presumably a following vowel also requiring a high tongue back (a high, back vowel) might compromise the perceptibility of that frication as a consequence of vowel anticipation (i.e. the listener factors out the acoustic effect of the high tongue back as simply the result of the following vowel and not intended in the consonant). This might result in the loss of the velar fricative element, yielding simply a labial-velar approximant /w/. An insurmountable problem arises from this account: it does not explain the occurrence of /hw/ in all the other forms in Spence where there is a following high, back vowel (*who*, *whose*, *whom*, *whole*, *whoop*), and also does not explain why these forms have the /h/ pronunciation (with labial dissimilation) in Johnston. The alternative, account (2) with a velarized labial fricative, performs much better. If /hw/ had labial and not velar frication in the varieties of Spence and Johnston, the loss of the initial fricative element in this word, but not the others, can easily be accounted for by positing a non-adjacent dissimilation in labial frication due to the final labiodental fricative /f/: /hw....f/ > /w...f/. Note that it is frication, not labiality that is dissimilating on this account, thus yielding a correct prediction. Of course, this analysis opens up a further area of phonetic/phonological investigation as to the variation in the phonetic realisation of /hw/. Why, for example, should Spence in Newcastle and Johnston in London pattern together?

4.7.3 Unexpected /hw/ words in Kenrick and Perry

As noted in Section 4.6 above, both Kenrick and Perry unexpectedly have words with /hw/, in contrast to the usual /w/ in these authors. Identifying these words allows us to contemplate a phonological pattern.

(3) Kenrick's unexpected /hw/ words
  a. whelm, wherry, whether
  b. whisk, whiskers, whisper, whistle
  c. whitlow, Whitsuntide, whiz
(4) Perry's unexpected /hw/ words
  a. whelm
  b. whisk, whiskers, whisper (but NOT in whistle)

The clearest patterns are firstly that the following vowel is always front /ɛ/ or /ɪ/ (possibly enhancing velar frication with a raised tongue front), and secondly that there is a group of forms with /s/ following the vowel, possibly with a tighter restriction for Perry that the /s/ has to be in the same syllable (thus /hwɪsk/ *whisk*, but /wɪsəl/ *whistle*). No clear phonological explanations are forthcoming (e.g. /hw/ does not occur before all front vowels, or in all forms with /ɛ, ɪ/), but these patterns could form the starting-point for a more detailed investigation. The non-phonological explanation of onomatopoeia/symbolism for a number of these words is posited in Section 4.6.2, but this cannot account for all of the forms.

4.7.4 Stress and compounds

The final five words in the data list comprise forms in which 'wh' occurs word-internally, not initially: *somewhere*, *somewhat*, *overwhelm*, *nowhere*, and *elsewhere*. Not all of the authors list all of the words, but those that have initial /hw/ (Johnston, Spence, Sheridan, Walker, and S. Jones) also consistently have internal /hw/. However, Perry, the only author listing these words who usually has /w/, has pronunciations with /w/ for *somewhere*, *somewhat* and *nowhere*, /hw/ for *overwhelm* and *elsewhere*. Recall that Perry has some unexpected /hw/ words, which show some indistinct patterns Section 4.7.3). However, in this instance, the pattern is unambiguous: /hw/ is licensed internally in the onset of a primarily stressed syllable, whereas /w/ occurs as the unstressed-syllable onset variant. The stress markers are clearly indicated by Perry himself to corroborate this pattern.

(5) Stress (acute accent) and the distribution of word-internal /hw ~ w/ in Perry
  a. Onset of internal stressed syllable: /hw
     overwhélm, elsewhére
  b. Onset of internal unstressed syllable: /w/
     sómewhere, sómewhat, nówhere

This distribution is exactly paralleled by the distribution of internal /h/ in present-day English: /h/ is absent in the onset of an unstressed syllable (*véhicle* with no /h/ pronounced), but present in the onset of a stressed one (*vehícular*).

5. Conclusion

In this pilot study, we have repeatedly found that by systematically collating the different types of direct evidence afforded by the eighteenth-century pronouncing dictionaries (sounds and stress, contemporary commentary, geographical and chronological spread), and incorporating other types of evidence into our analyses (phonetics, phonology, etymology, typology, frequency), we have been able to posit accounts for many of the patterns which such an orderly approach to the data has allowed us to ascertain. Some of the analyses are more robust than others, but all of them at the very least present a starting-point for further investigation. This could progress in two ways: either by further examining the phonetics, phonology, etc. of the problem, or by incorporating much more eighteenth-century data in order to identify much more robust patterns and potential sources of explanation. It is hoped that the proposed database, incorporating evidence from a wider range of sources providing evidence for eighteenth-century phonology, will allow us to develop this latter approach so as to place the former on a much firmer footing.

Notes

[1] This project was funded by the Rapid Response Fund of the Faculty of Arts, University of Sheffield, and the research assistant was Christine Wallis.

[2] M. Jones (2008) finds that /ʍ/ has a wider lip aperture than /w/, which would be opposite relationship expected between approximant /w/ and fricative /ʍ/. A possible explanation is that the lip aperture used to execute Scottish /w/ is too small to produce audible labial frication.

[3] The cluster analysis is in line with that of Vachek (1954), who presents other arguments regarding sonorant-cluster simplification in its favour, and Hickey (1984, 2007: 319-320), who presents arguments from sonority.

Sources

A Representative Corpus of Historical English Registers (ARCHER). Available online at http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/

A Corpus of Late Eighteenth-Century Prose. Available online at http://personalpages.manchester.ac.uk/staff/david.denison/late18c

Corpus of Early English Correspondence Extension (CEECE). Available online at http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/ceece.html

Corpus of Historical American English (COHA). Available online at http://corpus.byu.edu/coha/

Corpus of Late Modern English Texts (extended version) (CLMETEV). Available online at http://perswww.kuleuven.be/~u0044428/

Diachronic Electronic Corpus of Tyneside English (DECTE). Available online at http://research.ncl.ac.uk/decte/

Eighteenth-Century Collections Online (ECCO). Available online at http://gale.cengage.co.uk/product-highlights/history/eighteenth-century-collections-online.aspx

Eighteenth-Century English Grammars (ECEG). Available online at http://www.llc.manchester.ac.uk/research/projects/eceg/ecegdatabase/

IvIE (Intonation in the British Isles). Available online at www.phon.ox.ac.uk/IViE/

Literature On Line (LION). Available online at http://lion.chadwyck.co.uk

Network of Eighteenth-Century English Texts (NEET)

Old Bailey Corpus. Available online at http://www.uni-giessen.de/oldbaileycorpus/

Old Bailey Online. Available online at http://www.oldbaileyonline.org/

Phonologie d'Anglais Contemporain (PAC). Available online at http://w3.pac.univ-tlse2.fr/

Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE). Available online at http://www.ling.upenn.edu/hist-corpora/

References

Abercrombie, D. (1981), 'Extending the Roman alphabet: some orthographic experiments of the past four centuries', in: R.E. Asher and E. Henderson (eds) *Towards a History of Phonetics*. Edinburgh: Edinburgh University Press. 207-224.

Anderson, W. and J. Corbett (2009), *Exploring English with Online Corpora.* Basingstoke: Palgrave Macmillan.

Beal, J. C. (1999), *English Pronunciation in the Eighteenth Century: Thomas Spence's 'Grand Repository of the English Language' (1775).* Oxford: Clarendon Press.

Beal, J. C. (2004), *English in Modern Times 1700-1945*. London: Arnold.

Beal, J. C. (2007), 'To explain the present: 18th and 19th-century antecedents of 21st-century levelling and diffusion', in: J.L. Bueno Alonso, D. González Álvarez, J. Pérez-Guerra and E. Rama Martínez (eds) *'Of Varying Language and Opposing Creed': New Insights into Late Modern English.* Bern: Peter Lang. 25-46.

Beal, J. C. (2012a), 'Can't see the wood for the trees? Corpora and the study of Late Modern English', in: M. Markus, Y. Iyeiri, R. Heuberger and E.

Chamson (eds) *Middle and Modern English Corpus Linguistics: A Multidimensional Approach*. Amsterdam: Benjamins. 13-29.

Beal, J. C. (2012b), 'Evidence from Sources after 1500', in: T. Nevalainen and E.C. Traugott (eds) *Handbook on the History of English: Rethinking Approaches to the History of English*. Oxford: OUP. 63-77.

Benzie, W. (1972), *The Dublin Orator*. Menston: Scolar Press.

Buchanan, J. (1757), *Linguae Britannicae Vera Pronuntiatio*. London: A. Millar.

Burn, J. (1786), *A Pronouncing Dictionary of the English Language*. 2nd edn. Glasgow: Alex Adam for the Author and James Duncan.

Cooper, C. (1687), *The English Teacher*. London.

Denison, D. (1998), 'Syntax', in S. Romaine (ed.) *The Cambridge History of the English Language,* vol. IV, 1776-1997. 92-329.

De Smet, H. (2005), 'A corpus of Late Modern English', *ICAME Journal* 29: 69-82.

Dobson, E. J. (1957), *English Pronunciation 1500-1700. Vol. 2, Phonology*. London: OUP.

Douglas, S. [1779] (1991), *A Treatise on the Provincial Dialect of Scotland*. ed. by Charles Jones. Edinburgh: Edinburgh University Press.

Fitzmaurice, S. (2007), 'Questions of standardization and representativeness in the development of social network-based corpora: the story of the Network of Eighteenth-century English Texts', in: J. C. Beal, K. P. Corrigan and H. L. Moisl (eds) *Creating and Digitizing Language Corpora. Vol. 2: Diachronic Databases*. Basingstoke: Palgrave Macmillan. 49-81.

Fritz, C. (2007), *From Early English in Australia to Australian English 1788-1900*. Frankfurt: Peter Lang.

Görlach, M. (2001), *Eighteenth-Century English*. Heidelberg: Winter.

Hickey, R. (1984), 'Syllable onsets in Irish English', *Word* 35: 67-74.

Hickey, R. (2004), *A Sound Atlas of Irish English*. Berlin: Mouton de Gruyter.

Hickey, R. (2007), *Irish English: History and Present-day Forms*. Cambridge: CUP.

Hickey, R. (2010), *Eighteenth-Century English.* Cambridge: CUP.

Hodges, R. (1644), *The English Primrose.*

Johnston, W. (1764), *Pronouncing and Spelling Dictionary*. London: the Author.

Johnston, W. (1772), *Pronouncing and Spelling Dictionary*. 2nd edn. London: the Author.

Jones, C. (1989), *A History of English Phonology.* London: Longman.

Jones, C. (2006), *English Pronunciation in the Eighteenth and Nineteenth Centuries.* Basingstoke: Palgrave Macmillan.

Jones, M. (2008), 'What? Whence? And whither? A phonetic analysis of Scottish English "wh"'. Paper presented at the BAAP Colloquium, University of Sheffield, Tuesday 1st April 2008.

Jones, S. (1797), *Sheridan Improved: A General Pronouncing and Explanatory Dictionary of the English Language.* London: Verner and Hood, J. Cushell, Ogilvie and Son and Lackington, Allen and Co.

Kenrick, W. (1773), *A New Dictionary of the English Language*. London: John and Francis Rivington, William Johnston et al.

Kökeritz, H. (1953), *Shakespeare's Pronunciation.* New Haven: Yale University Press.

Kytö, M., M. Rydén, M. and E. Smitterberg (2006), *Nineteenth-century English: Stability and Change.* Cambridge: CUP.

Ladefoged, P. and I. Maddieson (1996), *The Sounds of the World's Languages*. Oxford: Blackwell.

Local, J. (1983), Making a transcription: the evolution of A. J. Ellis's Palaeotype. *Journal of the International Phonetic Association* 13: 2-12.

MacMahon, M. K. C. (1998), 'Phonology', in: S. Romaine (ed) *The Cambridge History of the English Language,* vol. IV, 1776-1997. Cambridge: CUP. 373-535.

Maguire, W. (2012), 'Mapping *The Existing Phonology of English Dialects*', *Dialectologica et Geolinguistica* 20: 84-107.

OED = *The Oxford English Dictionary*. 2nd edn. OED Online. Oxford. (1989). Available online at http://dictionary.oed.com/.

Perry, W. (1775), *The Royal Standard English Dictionary.* Edinburgh*:* David Willison for the Author.

Rissannen, M., M. Kytö, L. Kahlas-Tarkka, M. Kilpiö, S. Nevanlinna, I. Taavitsainen, T. Nevalainen and H. Raumolin-Brunberg (1991), *The Helsinki Corpus of English Texts*. Department of English, University of Helsinki.

Robinson, R. (1617), *The Art of Pronuntiation.* London: Nicolas Okes.

Sheridan, T. (1761), *A Dissertation on the Causes of the Difficulties which Occur in Learning the English Tongue*. London: R. and J. Dodsley.

Sheridan, T. (1780), *A General Dictionary of the English Language.* London: W. Strahan.

Smitterberg, E. (2005), *The Progressive in 19th-century English: a Process of Integration*. Amsterdam: Rodopi.

Spence, T. (1775), *The Grand Repository of the English Language*. Newcastle: Thomas Saint.

Tieken-Boon van Ostade, I. (2009), *An Introduction to Late Modern English*. Edinburgh: Edinburgh University Press.

Vachek, J. (1954), 'On the phonetic and phonemic problems of the southern English WH-sounds'. *Zeitschrift für Phonetik und Allgemeine Sprachwissenschaft* 8: 165-194.

The computer as research assistant: a new approach to variable patterns in corpus data

Gregory Garretson and Henrik Kaatari

Uppsala University

Abstract

This article advocates a particular type of semi-automated approach to working with corpus data termed "shared evaluation", the central idea of which is that the computer takes over more of the work of sorting and classifying the data, while a subsequent pass by a human coder ensures the ultimate accuracy of the data selection and classification. The article begins with a discussion of the traditional approach to corpus data and the tools that are currently available. It then describes the shared evaluation approach and compares this to a typical concordancer-based approach. The article goes on to present SVEP, a computer program developed by the authors to implement this approach and offered freely to other researchers, describing the most significant aspects of the program and its use. A case study involving adjective complementation is then presented, including examples of how SVEP was used in the study and an evaluation of the accuracy the program achieved. The article ends with a discussion of the advantages and disadvantages of SVEP in particular (and some ways the program might be improved) and of semi-automated approaches such as shared evaluation in general.

1. Introduction: the current state of the art

Traditionally, corpus linguists have worked primarily with standalone concordancers such as WordSmith Tools (Scott 2012), AntConc (Anthony 2011), and MonoConc Pro (Barlow 2000). These tools target a spectrum of users ranging from beginners to experts. As a result, they manifest a trade-off between providing a user-friendly interface and providing a powerful and flexible set of search capabilities. The flexibility of a program is determined by how many types of searches may be performed. For example, both AntConc and MonoConc Pro support regular expressions, thereby providing a very flexible search syntax (see Reppen 2001 for a review of MonoConc Pro and WordSmith Tools, and Weichmann and Fuhs 2006 for a review of ten different corpus search tools). In addition, many of these tools support batch searches, in which a user can perform several different searches at the same time. This is especially useful when searching for constructions that can appear in many different forms.

Several online tools also offer powerful search capabilities. The addition of the CQP (Corpus Query Processor) syntax to the BNC*web* interface to the *British National Corpus* (hereafter BNC; Burnard 2007) greatly improved the flexibility with which it can be used (see Hoffmann and Evert 2006; Hoffmann et al. 2008).[1] Previously, BNC*web* was based on the SARA server software and suffered from the limitation that searches needed to be based on a specific lexical item, making it impossible to search for grammatical patterns based solely on POS tags.[2] With the introduction of the CQP syntax, BNC*web* now offers highly complex searches: for one thing, searches can combine different layers of annotation such as part of speech, lemma and word. For another, the CQP syntax also allows for regular expressions, including the | operator, meaning "or", thereby enabling different patterns to be represented by one search string. In addition, BNC*web* allows the user to specify intervals of different ranges between different tags, as well as whether a particular tag should occur at the beginning or at the end of a sentence (see Hoffmann et al. 2008). As a consequence, for those working with the BNC, BNC*web* is now one of the most powerful search interfaces available.

However, a problem with search tools that permit searches allowing for a great deal of variation is that many of the matches returned will be "false positives", or undesirable hits. Another problem is that matches representing different patterns will all be mixed in together in the search results. This complicates greatly the tasks of sorting and classifying the data.

In our research on complex grammatical patterns, we felt that it would be desirable to be able to differentiate matches with different amounts of material intervening between the fixed portions of a search pattern. For example, when specifying an interval of 0-10 words, it would be helpful in evaluating the matches to differentiate between tokens with one word, a few words, several words, etc., in the interval, as illustrated by the examples of adjectival complementation in (1), where the node (i.e., the central element of the search, in this case an adjective) is shown in italics, the complement is shown in bold, and the interval material is underlined.[3]

(1) a. Over the next few days it became *clear* **that Jake was avoiding her**. (JXS)

b. It was *inevitable* therefore **that she should have looked for a career in motor racing** – at least that was what she told herself. (HGM)

c. It was *clear* by the late 1950s **that TV, like radio before it, would provide the 'main facts' and the latest news, and that people wanted to follow up the details in a newspaper**. (CRY)

d. If that goes up I would say it is highly *probable* – in fact think I should use the word inevitable – **that both the thermal plume and the temporarily quiescent earthquake zone along the tectonic fault would be reactivated**. (CKC)

Especially when an interval has the potential to include a large amount of material, it is desirable to differentiate hits with a small amount of material (which are more likely to be "good matches") from those with a large amount of material (which are more likely to be "bad matches"). However, with existing tools, this would require performing many different searches. A grammatical pattern with four separate intervals, for example, could require dozens of independent searches. One of the central features of the program SVEP described in Section 3 is that it classifies search matches according to the amount of "interval" material found in them after just one round of searching.

Another limitation that linguists must contend with is the fact that different corpora are associated with different tools. Researchers working with the BNC may benefit from the BNC*web* interface just described or the also powerful BYU-BNC interface (Davies 2004); meanwhile, those working with the ICE corpus family may make use of the powerful ICECUP program (Nelson et al. 2002). This program takes advantage of the fact that the ICE corpora are not only POS-tagged but also parsed, which naturally facilitates all manner of syntactic investigations. Nevertheless, programs such as ICECUP do not assist the user in the classification of the tokens, or in differentiating between "good" and "bad" matches.[4] And for many other corpora, there is no dedicated program for performing searches, so researchers must use a general-purpose tool.

In general, it is impossible for a single tool designed for a wide range of users to be able to incorporate all imaginable search features. For this reason, customized programs are sometimes necessary to efficiently extract complex grammatical patterns from corpora (see Hoffmann and Evert 2006: 192). In some cases, it is feasible for a research team to develop its own tools, but a more attractive option is to make use of tools developed by others, with minor modifications (see Garretson 2008). We believe that the approach described below may prove useful for a range of different studies, and so we are making our program SVEP freely available for the use of other researchers. Below, in Section 2, we describe the approach; in Section 3, we present the tool developed, and in Section 4, we present a case study demonstrating the approach. Finally, in Section 5, we critically evaluate the software and the approach.

2. The "shared evaluation" approach

As the above discussion suggests, a central issue in corpus linguistics is that the tools we have available govern the types of studies we may perform. Of course, this works in two ways: on the one hand, the development of new tools opens up new horizons for empirically based linguistic inquiry. On the other hand, the limitations of our tools also tend to curb our inclination to pursue new types of studies. Ideally, we would begin by imagining the study we wished to conduct, and then find or, if necessary, develop the appropriate tools. What we present here is just such a case, in which we envisioned a new methodology for working with variable patterns in corpus data, found that there were no existing tools that would make it feasible, and therefore developed a tool to implement the methodology.

We refer to the approach presented here as "shared evaluation", because it involves having the computer take on a more significant role in the evaluation of the data – in a way, the computer can be seen as a research assistant, taking over a great deal of the time-consuming labor. However, as with any research assistant, the ultimate responsibility for the analysis lies with the principal investigator, and so a certain degree of supervision is called for. Here, this takes the form of two rounds of evaluation: one by the computer, and one by the researcher.

2.1 Characteristics of the approach

The shared evaluation approach has eight characteristics; these are listed in Table 1. They need not all coincide, but we believe that the combination of all eight yields distinct advantages for the quality and facility of the research.

Table 1. Characteristics of the shared evaluation approach

Characteristic	**Brief explanation**
Automated first round of evaluation	Computer makes first pass at classifying tokens
Manual second round of evaluation	Researcher checks computer's classification
Token scoring	Each token receives scores based on heuristics
Batch searching	All searches are performed simultaneously
Zero duplication of data	Each token is recorded in only one place
Inclusion of metadata	Relevant metadata is output with tokens
Maximal recall	System captures as many tokens as possible
High ultimate level of precision	Data checked carefully to ensure accurate coding

The two central characteristics concern the classification of the data, which in the case of the study presented below means assigning each adjective found to one of fifteen different patterns of complementation, or discarding it. In this approach, the computer performs a first, tentative round of classification based on a set of heuristics, and then the researcher goes through the computer's choices and corrects them as necessary. We find that this method yields significant advantages in terms of both the number of tokens that may be included and the ultimate accuracy of the classification of the data.

The computer classifies the data by assigning scores to each token. When there are several different patterns that a token might match, the computer should assign it a score for each pattern; the token is then matched to the pattern for which it received the highest score. The current implementation of this approach, SVEP, uses a rather rudimentary basis for assigning the scores: a count of the number of words intervening between elements in the patterns. However, in principle the scoring of the tokens could (and probably should) be more complex and more linguistically informed.

Another crucial aspect of the process is that the computer performs all of the searches in parallel; that is, for every potential token, it compares it to every one of the patterns. This batch searching enables each token to be assigned to the optimal pattern. Another advantage of this is that there is zero duplication of data – each token is assigned to one and only one pattern, meaning that the researcher will not have to evaluate the same token more than once, which becomes especially important when thousands of tokens are involved.

It is frequently desirable to record metadata for each individual token – what file it came from, what its sentence ID is, what genre the text belongs to, who the speaker is (in the case of spoken data), etc. In our approach, the computer outputs all of the relevant metadata with each token, together with the text of the token.

A central aim of the shared evaluation approach is to enable larger datasets; in fact, this approach makes it possible to set the goal of capturing every single token of the phenomenon in question in the corpus – what we refer to as "maximal recall". There is a difference between randomly selecting 10,000 tokens from the entire BNC and retrieving *all* 10,000 tokens from a well-defined subcorpus; the latter approach allows us to have total accountability for the phenomenon in question, at least where a particular sub-corpus is concerned. To support this goal, the program SVEP saves all tokens, even those that have been rejected, to files where they may be inspected if desired.

Of course, in a situation in which total accountability is *not* required, this approach works excellently for finding a number of high-quality tokens. All that is required is to sort the tokens by score; the tokens at the top of the list will usually be very good matches to the pattern (such as 1a above), while those that are more marginal (or even misassigned) will fall to the bottom.

The final characteristic of this approach, thanks to the combination of both automated and manual passes of evaluating the data, is a high level of precision where the final classification of the tokens is concerned. The automated first round will have only moderate precision, but it creates a good starting point for the manual second round, which will improve the classification, resulting in few or no misclassified tokens; see Section 5 for more discussion.

2.2 Division of labor

The shared evaluation approach is based in large measure on the idea that the computer should take over some of the work of evaluating and classifying tokens from the researcher. It therefore shifts the responsibility for certain tasks in the process of extracting corpus data. To illustrate this, let us compare shared evaluation, instantiated here using the program SVEP (described in the next section) to a more well-known approach, using a tool such as WordSmith Tools or AntConc, which we will call the "concordancer approach".

Table 2. Division of labor in two approaches to corpus searches

	Concordancer Approach	**Shared Evaluation Approach**
1. Determine the targets	Researcher	Researcher
2. Formulate the queries	Researcher	Researcher
3. Search for and export the data	Computer + Researcher	Computer
4. Record metadata	Researcher	Computer
5. Eliminate duplicate tokens	Researcher	Computer
6. Discard irrelevant tokens	Researcher	Computer + Researcher
7. Classify the tokens	Researcher	Computer + Researcher
8. Analyse the data	Researcher	Researcher

Table 2 lists eight steps that are likely to be involved in the process of searching for, extracting, and classifying corpus data, and shows, for both approaches, who does most of the work in each step. The first step, included for the sake of completeness, is to decide exactly what patterns the phenomenon under study can take; the second step is to set up the program to search for those patterns by formulating search queries and otherwise tweaking the settings. The third step is to perform the search and save the results. Here we encounter the first difference between the methods: when performing several searches using a concordancer, the researcher has to set up and execute each search separately, and then save the results. With SVEP, once step 2 is complete, the computer performs all of the searches as a batch, saving the results for each search to a separate file. This means that the entire process can be repeated easily if further tweaks to the search patterns are called for.

The fourth step, recording metadata for each token such as the file name, text genre, sentence ID and so on is not made particularly easy by most concordancers; by contrast, SVEP gathers the relevant metadata for each token and outputs it with the token. Of course, the program expects a certain corpus format (currently that of the BNC); to extract metadata from differing corpora will require tweaking of the program.

The fifth step refers to the fact that similar searches may very well yield overlapping results, such that some tokens are found in more than one file, making it necessary to discard the duplicates. However, this does not happen in the shared evaluation approach, because of the way the data is managed: each token is compared to each pattern and then assigned to exactly one of them. It is also necessary to discard spurious tokens (false positives); the program SVEP facilitates this sixth step by exporting low-scoring tokens (see below) – those which are almost certainly bad matches – to an “OTHER” file.

The seventh step, classifying the tokens, can be rather difficult, depending on the phenomenon under study and the number of patterns. While this is not terribly difficult for a phenomenon such as the possessive alternation (compare *this author’s work* and *the work of this author*), it can be rather tricky in a case such as the adjective complementation study described below, with fifteen different patterns to choose among. Luckily, SVEP not only assigns each token to a pattern but also records its score, showing how closely the token appears to match that pattern; these scores may then be used as a guide in checking the program’s classification of the tokens.

The final step is of course the full analysis of the data; what form this takes will vary from study to study, and while it is certainly possible for this to be carried out in part by the computer (see Garretson and O’Connor 2007), that is not an integral part of the approach advocated here.

3. The program SVEP

The program we developed to implement the method described above is called SVEP, standing for System for Variable Extraction of Patterns.[5] It is written in the programming language Perl; the Perl interpreter comes pre-installed on many computers and otherwise can easily be installed for free. The program is designed to be parsimonious with resources, so it will run on a modest computer, and to be flexible but no-frills; it does not have a graphical user interface (GUI), but rather is run at the command line (i.e., at the command prompt, in a terminal window). Running the program is trivially easy; however, before it can be run, it must be configured for the corpus and the phenomenon under study. This is done by creating (or modifying existing copies of) various input files, which are plain text files written using special conventions that the program is designed to interpret. This is explained in greater detail below.

The main operation that the program performs is to search each sentence in the corpus for text matching a set of patterns supplied by the user. The point of departure is the *node*, which is the one element that must be present in any token. In the case study presented below, the node is defined as any word tagged as an adjective. The program compares each token found (e.g., every instance of an adjective in the corpus) to each of the patterns, assigning the token a score for each one, to show how closely that token appears to match that pattern. Once a given token has been compared to all possible patterns, the program "makes a call" and matches the token to the pattern for which it has the highest score; the token is then output to the results file corresponding to that pattern.

Currently, the scoring system used by SVEP is fairly simplistic: it counts the words found in each "interval" (see below) in the pattern in question, and gives more points to tokens with less intervening material, on the assumption that those tokens are more likely to be a good match to the pattern. High-scoring tokens are usually a good match to the pattern to which they are assigned. Frequently, however, lower-scoring tokens will be misclassified by the program and must be reclassified by the researcher. How often this happens will depend on the complexity and distinctness of the patterns in use.

As mentioned, before the program is run, it must be configured by creating a number of input files. For convenience, the program is distributed with samples of all of these files, and they are explained in detail in the user manual. However, to give a sense of how easy (or difficult) it is in practice to use this implementation of the shared evaluation approach, the input files and output files will be described briefly in the following sections.

3.1 The pattern file

The central input to SVEP is the pattern file, which is a simple plain-text file that defines the syntactic patterns that the program is to search for. Example (2) shows two of the patterns from the pattern file used in the study on adjective complementation discussed in Section 4. They are both designed to find instances of extraposition (e.g., *it is unfortunate that they refused*), with the second pattern covering cases of *that*-omission.

(2)
```
01 THAT_EXT:    IT [0:1] VERB [0:5] NODE [0:5] THAT [1:10] VERB
11 THAT_EXT_NT: IT [0:1] VERB [0:5] NODE            [1:4!] VERB
```

Each line of the file represents one pattern and consists of three "fields". The first field gives the pattern a number (e.g., 01); this is the rank of the pattern vis-à-vis the other patterns. In the case of a tie in score between two patterns, the program will assign the token to the pattern with the higher rank (i.e., the lower number). Using these numbers rather than simply relying on the ordering of the patterns in the file allows the user to order the patterns in a way that makes the relationship between them visually clear (see example 15 in Section 4.1). The second field in each line is the name of the pattern (e.g., THAT_EXT). This will be used to name all of the results files corresponding to this pattern (see Section 3.5).

The third field in each line is the pattern itself. This is a schematic representation of the sequence of things the program is meant to search for. It consists of a combination of "elements" and "intervals". The elements correspond to textual material; each one is represented by a code (e.g., VERB) that is defined in the regular expression file, as described below. The advantage of removing the nitty-gritty of the regular expressions to another file is that this allows the patterns shown here to be comparatively clear and uncluttered, which facilitates both designing and interpreting them.

The intervals, written as brackets with numbers (e.g., [0:5]), represent places between elements where optional material may occur; these are central in the calculation of the score for each token. In each interval, the number before the colon represents the lowest number of words that may occur here.[6] The second number is the maximum number of words that may appear here for a candidate to be considered a "good match". Essentially, the more words occur in an interval, the lower the score for that token. This is based on the observation that, given two tokens that are both possible matches to a pattern, the one with more intervening material is more likely to turn out *not* to be a valid match. The consequent lowering of the score for that token indicates a lowered confidence that the token does indeed match that pattern and increases the likelihood that it will ultimately be assigned to another pattern. All of the sentences in (1) above turn out to be valid examples of the THAT_EXT pattern, but they vary in the score they receive, with (1a) receiving the highest score and (1d) receiving the lowest score.

One of the advantages of the method used here is that tokens with material intervening at several different points in the pattern – compare example (1d) to the first pattern in (2) – are not missed by the search, as would be likely given a more conservative approach. Rather, they are included but given a lower score. The scores thus serve two purposes. First, they are used by the program to select the pattern to which each token will be assigned. Second, they may be used by the researcher when examining the tokens to figure out which ones are properly classified and which ones need to be reclassified. Sorting a set of tokens by their score frequently results in a set of very good matches at the top, and then an increasing preponderance of unusual or misclassified tokens toward the bottom.

3.2 The filter file

The filter file uses exactly the same format as the pattern file, and the contents are used in the same way, but for a different purpose. The filter file presents patterns (such as that in example 3) to be *excluded* from the results in a filtering sweep that the program performs *before* it begins matching the corpus against the pattern file.

(3)
```
02 NODE_NOUN:  NODE PUNCT NOUN
```

This is useful in those cases where it is easier to define undesirable sequences than it is to find a way to exclude such tokens via the patterns in the pattern file (or the regular expressions in the regular expression file). For example, when looking for adjectives with complements (e.g., *certain to occur*), it is convenient to first rule out all attributive adjectives (e.g., *a certain number*), since these will never have a complement.[7] Excluding these in the filtering sweep allows the patterns used in the main matching sweep to be kept simpler. Example (3) shows a pattern used to filter out attributive adjectives in the study discussed below.

3.3 The regular expression file

The regular expression file is the only part of the procedure that requires a measure of advanced technical knowledge. Its purpose is to define the codes used in the pattern file and the filter file in terms that the computer will understand. Essentially, this means defining them using Perl regular expressions.

Regular expressions are a specialized pattern-matching language that is extremely common in contexts where complex textual searches are performed: in programming languages such as Perl, Python, Java, etc., as well as in concordancers and other tools allowing advanced text searches. The regular expressions used by Perl, a language known for its excellent text-search capabilities, are very similar to those used in other contexts. Today most concordancers offer the *option* of using regular expressions; what is different about SVEP is that it *requires* their use.

The fundamental idea behind regular expressions is to define classes of items to match. For example, the symbol \w stands for “any letter, digit, or underscore character”, and the combination \w+ stands for one or more of these. This makes it possible to craft search terms that match a range of possible textual strings, rather than just one.

Example (4) shows the contents of the regular expression file used in the study presented in Section 4. Each line of the file consists of two fields: a code (followed by a colon), and the definition of that code.[8]

(4)

```
NODE:     <w\ c5="(AJ\w|AJ\w-\w\w\w|\w\w\w-AJ\w)" [^<]+ </w>
ANY:      .*?
END:      [^<]+ </w>
PUNCT:    (?:<c\ [^<]+ </c>)*
COMMA:    <c\ [^<]+ </c>
[#:       <s\ [^>]+ >
#]:       __PUNCT__ </s>
HIT:      <hit><w\ [^<]+ </w></hit>
THAT:     <w\ c5="(?:CJT|CJT-DT0|DT0-CJT)"       __END__
FOR:      <w\ c5="PRP"\ hw="for"                 __END__
IT:       <w\ c5="PNP"\ hw="it"                  __END__
NOUN:     <w\ c5="N\w\w"                         __END__
NP:       <w\ c5="(?:PNP|N\w\w)"                 __END__
INF:      <w\ c5="V(?:VB|\wI)"                   __END__
VERB:     <w\ c5="V\w\w(?:-V\w\w)?"              __END__
TO:       <w\ c5="TO0"                           __END__
ART:      <w\ c5="AT0"                           __END__
ADJ:      <w\ c5="(?:AJ0|AJ0-\w\w\w|\w\w\w-AJ0)" __END__
CONJ:     <w\ c5="CJC"                           __END__
```

The definition includes two types of content. For the most part, it consists of text using regular expressions to be matched against the content of the corpus files. The exact nature of this text will depend on how the corpus files are marked up; for example, here the XML version of the BNC is the corpus being searched, and so the regular expressions are designed to match text like that in (5) below – compare this to (1a).

(5)

```
<s n="3663"><w c5="PRP" hw="over" pos="PREP">Over </w><w c5="AT0"
hw="the" pos="ART">the </w><w c5="ORD" hw="next" pos="ADJ">next </w><w
c5="DT0" hw="few" pos="ADJ">few </w><w c5="NN2" hw="day"
pos="SUBST">days </w><w c5="PNP" hw="it" pos="PRON">it </w><w c5="VVD"
hw="become" pos="VERB">became </w><w c5="AJ0" hw="clear"
pos="ADJ">clear </w><w c5="CJT" hw="that" pos="CONJ">that </w><w
c5="NP0" hw="jake" pos="SUBST">Jake </w><w c5="VBD" hw="be"
pos="VERB">was </w><w c5="VVG" hw="avoid" pos="VERB">avoiding </w><w
c5="PNP" hw="she" pos="PRON">her</w><c c5="PUN">.</c></s> (JXS)
```

In addition, to increase the functionality and elegance of the pattern definitions, they may also include *codes that are defined in the same file*. The only requirement for doing so is that the code, when used, must be surrounded by double underscores (e.g., __END__). For example, in (4), the code END is defined as meaning "[^<]+ </w>"; this tells the program to carry on until the end of the word is found. Since this is a bit of regular expression material that is frequently needed when matching against the BNC, it is used in most of the other lines by inserting __END__. This helps to keep the definitions uncluttered and more readily interpretable at the same time that it minimizes the risk of typographical errors.

3.4 Other input files

SVEP makes use of four other input files apart from the three described above: a settings file, a stop word file, a corpus file, and a genre file. All but the settings file are optional. This section presents a very brief description of each of these input files.

The purpose of the settings file is to allow the user to tweak the program in various ways, such as modifying the scoring system, defining what counts as a token for the purposes of interval measurement (see note 6), and requesting tabs between words in the TXT output files (see below).

The stop word file simply includes a list of lexical items that are *not* to be matched to the node in any pattern. This is especially useful when certain words in the corpus have been tagged in a questionable manner. For example, in the BNC, it is quite common for the words *Tory* and *Labour* to be tagged as adjectives, but since they are not adjectives that allow complementation, it proved convenient to exclude them from consideration via the stop word file.

The corpus file allows the user to define a list of corpus files to be searched, in case it is not desirable to search the entire corpus. This is especially useful for testing purposes, when only a limited amount of output is needed.

Finally, the genre file was included in response to the fact that the genre of a BNC text is not recorded in its file header (by contrast to the domain, which is recorded). Especially since we are using a highly detailed, custom genre analysis of the BNC files (see Kaatari 2012), it proved necessary to supply the genre information independently. In theory, this mechanism could easily be adapted to supply whatever other metadata information one wished to include in the output files.

3.5 Program output

As mentioned above, SVEP outputs the search results, already classified, to a number of files, one per pattern. Since each token is assigned to exactly one pattern, each token is output to only one file, meaning that there is zero duplication of data. In order to allow total accountability of the data, the program can also be instructed to output each token that is matched to one of the filters (see Section 3.2) to a file named after that filter, so that all of the rejected tokens may be inspected if desired. This means that no data is ever fully discarded.

All of the tokens for a given pattern are output to an XML file named after that pattern. There is also an output file called "OTHER", which contains all tokens that were not filtered out but had no score greater than zero for any pattern. These tend to match one or more patterns very marginally. In addition, all the material is output in a tab-delimited plain-text format. These two versions are useful for different purposes. The XML file is suitable for further processing, using tools designed to work with XML formats (e.g., an XPATH search tool or an XSLT processor) or custom tools. The tab-delimited file is ideal for pasting into a spreadsheet (such as Microsoft Excel) for manual examination and coding. A convenient feature of the program is that the user may request that a number of words on either side of the node be separated with tabs, which means that they will fall into separate columns in the spreadsheet, thereby making it possible to sort the tokens, KWIC-style, by the first word to the left or right, etc.

SVEP outputs a certain amount of metadata with each token: a unique token ID, the ID of the sentence in which it was found, the ID of the text in which it was found, the genre of that text, the POS tag found on the node, the pattern to which it was matched, and the score it received for that pattern. In addition, the token is output with the entire s-unit in which it was found, so that it is relatively easy to evaluate the token in context. If further metadata were desired, the program could be tweaked to make this possible.

4. A case study: adjective complementation

SVEP was originally designed to solve a number of different problems in the data extraction stage of a study involving adjectives complemented by *that*- and *to*-clauses. Previous studies of adjective complementation have typically been confined to a predetermined set of lexemes (see, e.g., Mindt 2011, Van linden 2012), but the underlying idea of SVEP was to work around this lexical impasse and extract *all* adjectives that were found to be complemented by either of these two clause types. This proved to be a rather difficult task, since there are several different patterns of complementation, and there is a fair amount of variation within each pattern in terms of intervening material in the various intervals.

Adjective complementation by *that*- and *to*-clauses is represented by the following prototypical patterns (in which the adjective in question is shown in italics and the complement in bold; see also note 4):

(6) It was *inevitable* **that he should be nicknamed 'the Ferret'**, although seldom in his hearing. (CJF) (THAT_EXT)

(7) **That he should be nicknamed 'the Ferret'** was *inevitable*. (†) (THAT_PRE)

(8) I'm *happy* **that we are married**. (CEY) (THAT_POST)

(9) It is *difficult* **to test a potential cure when a disease is ill-defined**. (ARF) (TO_EXT)

(10) **To test a potential cure when a disease is ill-defined** is *difficult*. (†) (TO_PRE)

(11) Yet the authorities were *unable* **to silence the expression of political opposition**. (FB1) (TO_POST)

The examples in (6-11) represent the three basic constructions in which adjectives are complemented by *that*- and *to*-clauses. Here we find extraposed clauses (THAT_EXT and TO_EXT), such as (6) and (9), with non-referring *it* as matrix subject; pre-predicate clauses (THAT_PRE and TO_PRE), such as (7) and (10), in which the complement clause functions as matrix subject; and post-predicate clauses (THAT_POST and TO_POST), such as (8) and (11), with a referring pronoun (8) or a noun phrase (11) as matrix subject (terminology from Biber et al. 1999). However, these three constructions can be altered or expanded in a number of different ways, as illustrated in (12-14), thereby creating additional patterns that have to be accounted for.

(12) I'm *happy* [ø] **we are married**. (†) (THAT_POST_NT)
(13) It is *difficult* **<u>for scientists</u> to test a potential cure when a disease is ill-defined**. (†) (TO_EXT_FOR)
(14) *Unable* **to silence the expression of political opposition**, the authorities simply watched. (†) (TO_POST_NV)

This variability within the constructions is manifested in *that*-omission (NT, standing for "no *that*"), as in (12); the inclusion of *for*-subjects (FOR), as in (13); and the absence of a matrix verb (NV, standing for "no verb"), as in (14). When combining these features across the three constructions, we find a considerable number of different patterns. In fact, there are as many as 15 different patterns that have to be accounted for, as listed in Table 3. The way each pattern is defined for SVEP is shown below in (15).

Table 3. The 15 patterns listed by construction and complementation type

	Extraposed	**Pre-predicate**	**Post-predicate**
***that*-complementation**	THAT_EXT THAT_EXT_NT	THAT_PRE	THAT_POST THAT_POST_NV THAT_POST_NT THAT_POST_NV_NT
***to*-complementation**	TO_EXT TO_EXT_FOR	TO_PRE TO_PRE_FOR	TO_POST TO_POST_NV TO_POST_FOR TO_POST_NV_FOR

Not only are there many patterns to account for, but the structures of the patterns are very different from each other. Consider, for example, the difference between the extraposed – as in (6) and (9) – and the pre-predicate constructions – as in (7) and (10) – which vary considerably in terms of the position of the adjective and the position of the complementizer. Furthermore, the extraction of tokens with *that*-omission is problematic, since the complementizer itself is missing. Extraction of tokens with *that*-omission is thus typically achieved through searching for lexemes that are commonly found in this pattern and then identifying valid tokens through manual inspection (cf. Mindt 2011). For these reasons, pre-predicate clauses and *that*-omission are notoriously difficult and time-consuming to study. Below, we will show how SVEP deals with these difficulties; although manual inspection will always be essential in identifying these structures, SVEP manages to avoid being restricted to specific lexemes, which considerably increases the number of tokens found.

The remainder of this section demonstrates the use of the shared evaluation approach, specifically as implemented in SVEP, in a study of adjective complementation using the patterns detailed above. The corpus used in the study consists of a 3-million-word sample from the XML edition of the BNC including both spoken and written material; see Kaatari (2012) for a full description of the corpus. Sections 4.1 and 4.2 present the two most important input files, the pattern file and the filter file. The precision and recall of the program are evaluated in Section 4.3. The more general advantages and disadvantages of the methodology are discussed in Section 5.

4.1 The pattern file

The patterns listed above in Table 3 constitute the basis for the pattern file, which is shown in (15).[9] In this file, the patterns consist of elements and intervals (see Section 3.1); the elements have been given easily interpretable names to represent different parts of speech or lexemes. These codes are in turn defined in the regular expressions file (see Section 3.3).

(15)

```
13 THAT_PRE:         [# [0:1] THAT [0:3] NP [0:5] VERB [0:10] VERB [0:5] NODE [0:3]  #]
01 THAT_EXT:                                      IT   [0:1]  VERB [0:5] NODE [0:5]  THAT [1:10] VERB
11 THAT_EXT_NT:                                   IT   [0:1]  VERB [0:5] NODE             [1:4!] VERB
03 THAT_POST:                                                 VERB [0:5] NODE [0:5]  THAT [1:10] VERB
10 THAT_POST_NT:                                              VERB [0:5] NODE             [1:4!] VERB
06 THAT_POST_NV:                                              [#   [0:5] NODE [0:5!] THAT [1:10] VERB
15 THAT_POST_NV_NT:                                           [#   [0:5] NODE             [1:4!] VERB
14 TO_PRE_FOR:       [# [0:1] FOR  [1:5] TO [0:1] INF  [0:10] VERB [0:5] NODE [0:3]  #]
12 TO_PRE:           [# [0:1]            TO [0:1] INF  [0:10] VERB [0:5] NODE [0:3]  #]
05 TO_EXT_FOR:                                    IT   [0:1]  VERB [0:5] NODE [0:3]  FOR  [1:5]  TO [0:1] INF
02 TO_EXT:                                        IT   [0:1]  VERB [0:5] NODE             [0:5]  TO [0:1] INF
07 TO_POST_FOR:                                               VERB [0:5] NODE [0:3]  FOR  [1:5]  TO [0:1] INF
04 TO_POST:                                                   VERB [0:5] NODE             [0:5]  TO [0:1] INF
09 TO_POST_NV_FOR:                                            [#   [0:5] NODE [0:3]  FOR  [1:5]  TO [0:1] INF
08 TO_POST_NV:                                                [#   [0:5] NODE             [0:5!] TO [0:1] INF
```

As discussed in Section 3.1, the intervals, shown here as bracketed elements in the patterns, are crucial in determining the score a given token is assigned and, based on that score, which pattern the token is assigned to. To give an example, the THAT_EXT pattern in (16) includes four intervals with a maximum score of 20 points each.

(16) `THAT_EXT: IT [0:1] VERB [0:5] NODE [0:5] THAT [1:10] VERB`

In order for a token to receive the highest possible score, it must match the lower limit for each interval, with little or no intervening material. The token in (17) receives the maximum score of 80 points (20 × 4) when matched to the THAT_EXT pattern, since the lower limit for each interval is met. The elements matched by the program are underlined in (17).

(17) It is *certain* **that bees are very responsive to different tones of the human voice** […]. (G09)

As can be seen, in the first three intervals there is no intervening material, and the token thus matches the lower limit for these intervals, which is 0. In the fourth interval, the lower limit is 1, since the *that*-clause requires a subject, and this specification is also met by the token in (17), thus giving it the highest score possible for the THAT_EXT pattern.

The more material occurs within each interval, the lower the score will be, since the likelihood of the token being valid is reduced as the intervening material increases. This is illustrated by an invalid token in (18), in which the upper limit of the third interval ([0:5]) and the lower limit of the fourth interval ([1:10]) are violated, thus reducing the score to 40, indicating that this token is a questionable match to this pattern.

(18) We got up long before it was *necessary*, impeding all the sandwich-making and hard-boiling of eggs that was going on. (GVY)

A token is matched to the pattern for which it receives the highest score. In the case of (18), the score of 40 it receives for THAT_EXT is the highest score it receives for any of the 15 patterns, so it is assigned to that pattern, leaving it up to the researcher to discard the token during the manual round of evaluation.

Note that the limits for the intervals are not categorical; if there is more than one word intervening in an interval with the specification of [0:1], this simply means that the score for that interval is zero. However, a strict upper limit can be set to make the interval categorical, by including an exclamation mark – as in [0:1!] – thereby excluding such tokens. One other symbol that is used in the pattern file is a hash symbol coupled with a square bracket (e.g., [# or #]). This is used to represent the beginning or end of a sentence, respectively, and is useful when elements of a pattern have a tendency to occur sentence-initially or sentence-finally (e.g., the NV and PRE patterns in example 15).

The upper and lower limits used for the intervals in (15) were tested repeatedly using a test corpus of 100,000 words. Categorical intervals, such as those used for the patterns with *that*-omission (NT) and the patterns without a matrix verb (NV), were set on the basis of the maximum number of intervening elements found in the test data. This means that some valid tokens could in theory be missed; however, since SVEP can be instructed to save *all* tokens to files, even excluded tokens are still available for inspection. Section 4.3 below presents an analysis of the recall the program actually achieved.

4.2 The filter file

The filter file, as explained in Section 3.2, is designed to exclude tokens that are not of interest.[10] This feature is especially important when the patterns are centered on parts of speech that are very frequent. In this study, the filter file proved crucial in reducing the number of candidate tokens to a manageable number, since the corpus used contains around 226,000 adjectives, 192,000 of which were filtered out.

The filters are formulated in the same way as the patterns, using elements that are defined in the regular expressions file. The (slightly modified) contents of the filter file for the study on adjective complementation are given in (19).

(19)

```
01 ART_NODE:           ART NODE
02 NODE_NOUN:              NODE PUNCT NOUN
03 NODE_ADJ_NOUN:          NODE PUNCT ADJ PUNCT NOUN
04 NODE_CONJ_ADJ_NOUN:     NODE PUNCT CONJ ADJ PUNCT NOUN
```

The filters in (19) filter out attributive adjectives, since these cannot take complementation (see note 7). The filters serve this purpose in different ways by filtering out adjectives preceded by an article (01), adjectives followed by a noun (02), adjectives followed by another adjective and a noun (03), and adjectives followed by a conjunction, another adjective and a noun (04). These last two filters are illustrated in (20) and (21) below. Note also that filters 02–04 are designed to allow optional punctuation in various places (indicated by PUNCT).

(20) The more *important* sacred precincts were probably bounded by more substantial walls: the Gypsades Rhyton seems to show one of these. (CM9)

(21) This *lengthy* and expensive process is particularly unsuitable when the trustee suspects the assets may be sold off by the bankrupt or grabbed by local creditors. (AHT)

More precise figures on the tokens filtered out in this study are given in the next section.

4.3 Performance of the program

In this section, we evaluate the performance of SVEP in this study in terms of precision and recall. By precision, we mean its ability to correctly identify valid tokens (meaning ones that evince some form of adjective complementation) while rejecting invalid tokens (i.e., no false positives). By recall, we mean its ability to find all valid tokens without missing any (i.e., no false negatives). Beyond simply finding the tokens, the program also performs a pass of classification, the accuracy of which is also evaluated.

The starting point for the search was the set of all words in the corpus tagged as adjectives (including those tagged as *possibly* being adjectives; see below). In our three-million-word corpus, there are 225,978 such words in total; 192,088 of these adjectives (roughly 85%) were filtered out by the program, leaving 33,710 adjectives that were identified as potentially valid tokens. However, 1,569 of these received a score of zero on all patterns and were thus saved to the OTHER file (see Section 3.5); the likelihood of these being valid tokens is extremely small. This left 32,141 tokens that were identified as likely to be of interest. Table 4 shows how many tokens were assigned by the program to the fifteen patterns, with information on the accuracy of these assignments. The notation used in Table 4 indicates how many tokens were matched to the correct pattern (A), how many were correctly identified as valid but matched to an incorrect pattern (B), and how many were not valid tokens of adjective complementation but were not rejected by the program (C).

Table 4. Performance of the program in assigning tokens to the patterns

	A:	**B:**	**C:**		**(A+B)**
	Correct	**Incorrect**	**Invalid**		**÷**
Pattern	**pattern**	**pattern**	**token**	**Total**	**Total**
TO_EXT	1,448	374	882	2,704	0.67
TO_EXT_FOR	192	10	109	311	0.65
THAT_EXT	766	55	541	1,362	0.60
TO_POST	3,028	467	4,939	8,434	0.41
TO_POST_NV	129	24	381	534	0.29
THAT_POST	549	80	2,376	3,005	0.21
THAT_POST_NV	25	5	149	179	0.17
TO_POST_FOR	56	15	536	607	0.12
TO_PRE_FOR	1	12	164	177	0.07
THAT_PRE	7	129	1,750	1,886	0.07
THAT_EXT_NT	53	19	1,211	1,283	0.06
THAT_POST_NT	323	32	7,365	7,720	0.05
TO_PRE	18	43	1,855	1,916	0.03
TO_POST_NV_FOR	1	2	94	97	0.03
THAT_POST_NV_NT	40	4	1,882	1,926	0.02
Total	**6,636**	**1,271**	**24,234**	**32,141**	**0.25**

As seen in the bottom rightmost cell in Table 4, the overall precision of SVEP in identifying tokens of adjective complementation is not terribly high, at 25%. However, there are considerable differences between the different patterns. The extraposed patterns (EXT) at the top are extracted with fairly good precision (all of them above 60%), whereas the program does not perform as well on patterns with *that*-omission (NT). This is not surprising, given that these patterns lack the most salient indicator of complementation, viz. the complementizer *that*. The pre-predicate patterns (PRE) also exhibit a low rate of precision, due to the fact that these patterns not only are quite rare but also are easily confused with other structures. The program performs considerably better in terms of recall; this is by design, as discussed below.

In order to determine how many valid tokens of adjective complementation the program might have missed, we conducted a test: we extracted text from two randomly selected files from each of the 15 genres in the corpus, creating a 1% subsample of the corpus. We went through this subsample manually, looking for all instances of adjective complementation; 80 such instances were identified. These were then compared to the program's output. It was found that 78 of the 80 tokens had been positively identified by SVEP. The two remaining tokens were missed because they had been misclassified by the POS tagger CLAWS (Garside and Smith 1997); since they had not been tagged as adjectives in the corpus, they were missed by SVEP, which in this study relies crucially on POS tags.[11]

Interestingly, it was found that another of the 80 tokens had previously been discarded erroneously during the manual round of evaluation. This means that the only three instances of error found in this test were attributable not to SVEP, but to other factors. The rate of recall for the automated round of coding alone was thus

97.5% (due to the POS tagger errors), and the rate of recall for the automated and manual rounds together was 96.3% (due to tagger and human error). This means that roughly four percent of valid tokens in the data were missed.

We also evaluated the *precision* of the coding in the same subsample, to see whether any invalid tokens were incorrectly included. We examined all tokens from the subsample that had not been discarded during either the automated or the manual round and found that one invalid token had been retained. The rate of precision for the subsample is thus 98.7%, meaning that roughly one percent of the tokens included were in fact invalid.

These results, while good, illustrate that neither machines nor humans are perfect. If the goal is to maximize both recall and precision, the best way to achieve this is to use the computer to cast the widest net possible and go through the resulting data numerous times by hand. However, the benefits of an extreme version of this approach must be weighed against the time and effort required. We believe that the shared evaluation approach comes close to finding the “sweet spot” at which sufficient advantage is taken of the strengths of both computers and human analysts, yielding results that are very good both in terms of recall and in terms of precision.

It is worth pointing out that our goal for this particular study was in fact to find each and every token of adjective complementation in the corpus. In this quest for maximal recall, we instructed the program to cast a very wide net indeed, including not only words tagged clearly as adjectives (e.g., with the tag AJ0), but also those given “ambiguity tags” (e.g., AJ0-NN1), indicating that the POS tagger was not certain about the classification of the word. The same goes for other parts of speech as well as the important word *that* (which can at times be ambiguous for the tagger between a complementizer and a determiner). This increased significantly the number of invalid tokens matched and reduced the overall accuracy of the program, thereby increasing the amount of manual work required. If however, in another scenario, the requirement of finding *all* potential tokens were relaxed, it would be possible to tweak the pattern file, the filter file, and the definition of a node to reduce the number of tokens extracted. This would essentially increase the program’s precision at the expense of its recall, but it would also reduce radically the effort required in the manual inspection phase.

5. Evaluation and discussion

As seen above, the performance of SVEP alone (excluding the crucial manual stage) in our case study was characterized by high recall but low precision. This can be attributed, we believe, to three main factors. First, the algorithm for assigning scores to the various matches (recall that each token receives one score per pattern) is currently quite simplistic, employing merely a count of the words intervening between pattern elements. A more linguistically informed algorithm would almost certainly perform better. The second factor is the nature of the particular study presented above. The fact that we used fifteen different patterns, which in some cases differed only slightly, set the bar for matching a token to the correct pattern very high (random performance in the classification would give only 7% accuracy) and made it more likely that any given token would be accepted as valid. The third factor has to do with our goal for the study, as discussed above; we aimed to achieve 100% recall, meaning that the ultimate data set would include every single instance of adjective complementation in the three-million-word corpus, which is a tall order. As seen above, we came fairly close to this goal, but at the cost of lowering the precision – that is, by having the program include a number of marginal tokens which had to be evaluated by a human.

In our view, none of these factors speak against the shared evaluation approach itself, with which we have had an overall positive experience. In the next three sections, we will go over the advantages and disadvantages of SVEP (which is thus far the only implementation of the approach), mention some potential improvements to the program which might make it more useful in the future, and discuss the general advantages and disadvantages of semi-automated approaches such as shared evaluation.

5.1 Advantages and disadvantages of SVEP

The principal advantages of SVEP (as compared to other tools such as online or standalone concordancers), we believe, are as follows: it is a free, open-source program that may be modified as needed. Once configured via a pattern file, etc., it searches the corpus and outputs the results (together with a number of items of metadata) all at once, with no duplication in the output, meaning that each token is found in only one file. The files present the program's classification of the data, something that is not done by concordancers; this will vary in accuracy depending on the nature of the study, but each token is output with its score, which acts as an aid in evaluating the token. The fact that all of the searches are performed concurrently makes it very easy to tweak the patterns and the filters and then rerun the program. Importantly, the filtering sweep makes it possible to discard vast numbers of invalid tokens. The data is output both in XML format and in tab-delimited TXT format (with words around the node set off by tabs, for KWIC-style sorting), so that it may be further processed either by additional programs or manually. The program outputs each token marked up within the sentence (or s-unit) in which it is found, so that each token may be evaluated in context. In other words, SVEP goes beyond instantiating all of the characteristics of the shared evaluation approach presented in Section 2.1.

All in all, the program does a fairly good job of finding the data, organizing it, and presenting it for further evaluation, with a first round of classification performed before the researcher even sees it. This is the core idea of the shared evaluation process, and we have found it to be very helpful.

If we consider these advantages in light of the study presented above, we see how much utility they offer. Using other software to achieve the same results for the fifteen highly variable patterns of adjective complementation presented above would require dozens and dozens of searches, and these searches would in all likelihood have overlapping results, meaning that great quantities of data would have to be discarded manually. Furthermore, some of the patterns included in this study are notoriously hard to extract; even when finding these using a predetermined lexical list, researchers have to rely on time-consuming manual inspection. Even though the program's accuracy in classifying the patterns was often fairly low, this initial classification plus the inclusion of the tokens' scores facilitated the manual inspection stage considerably, with a token's score serving as an indicator of the likelihood (a) that the token was valid, and (b) that it had been correctly assigned to that pattern.

Turning now to the disadvantages of SVEP, these result primarily from the fact that it is a new tool that has not undergone a great deal of development. Many of them could be ameliorated by adding new features or making existing features more sophisticated; a few such potential improvements to the program will be discussed in the next section.

The principal disadvantage of the program is most likely the fact that it requires some understanding of regular expressions. While such knowledge is increasingly common among linguists – especially as they become increasingly familiar with languages such as R and sophisticated search interfaces such as that of BNC*web* – it is still by no means the norm. Yet it is the regular expressions that enable the highly complex matching that SVEP (or any other flexible search program) performs. In order to ameliorate this problem somewhat, we have abstracted the user input to the program into two layers: the regular expression file on one hand, and the pattern file (and the functionally similar filter file) on the other. While some understanding of regular expressions is required to manipulate the former, the latter is very simple to work with (compare examples 4 and 15 above). We also offer sample files with the program, which may be used as a starting point for other studies.

A second disadvantage of the current version of the program has to do with the scoring system by which the tokens are matched to patterns. As it stands, a token's score for a given pattern is calculated simply on the basis of the number of words found in each interval. While this works relatively well, it does not take the linguistic nature of the intervening material into account. A more sophisticated scoring system would be likely to perform better. It is also worth pointing out that a token that receives a lower score is not necessarily an aberrant instance of the pattern; it simply means that matching that token to that pattern is less likely to be correct, given the syntactic complexity of the sentence.

A more specific and less obvious disadvantage is that as the program is currently designed, each pattern must have a non-optional *node*; this node serves as the initial point of identification of a token. For example, in the study presented above, the point of departure was any word tagged as an adjective. Once an instance of an adjective had been identified, it could then be compared to each of the patterns. While a node may be defined in a variable way (for example, in the study just mentioned, it was defined as any adjective, including comparative forms, superlative forms, and any ambiguity tag that includes an adjectival tag), the fact remains that there must be a non-optional node in each pattern. This reduces, albeit by a small margin, the number of possible studies that may be conducted.

The program's reliance on POS tags might also be seen as a disadvantage; see Section 4.3 above for a discussion of how POS errors can negatively affect results. However, it should be pointed out that while the study described here makes crucial use of POS tags, this is not a requirement for SVEP – it would be possible to design patterns using only lexical information.

Currently, SVEP makes crucial use of sentences, or s-units, as they are marked up in the corpus. When the program matches tokens against the various patterns supplied by the user, it uses the entire s-unit as the context for the match. As mentioned above in Section 4.1, the user may even require that elements occur at or in proximity to sentence boundaries. However, the way the program presently works, it is not possible to match *across* sentence boundaries. This is typically not an obstacle for investigations of syntactic phenomena, but it could conceivably become an impediment to discourse-level studies.

Another result of the fact that SVEP has thus far been used exclusively with the BNC is that it is designed to extract metadata from that corpus (or another one of similar design). For example, the program is able to identify and record the ID of each s-unit, and a specially designed lookup function enables it to output with each token the genre of the text in which it occurs. When SVEP is used with other corpora, it will be necessary to make adjustments to the program code to ensure that the desired metadata is found correctly.

One final disadvantage is the fact that SVEP is a command-line tool, without the graphical user interface that most researchers have come to expect. While a GUI interface would doubtless prove comforting, the fact that the input to the program is overwhelmingly textual makes it questionable whether adding this feature would be worth the development time required.

5.2 Potential improvements to the program

Having detailed in the previous section several disadvantages of the current version of SVEP, we will briefly list here a number of potential improvements to future versions of the program. Whether these improvements will in fact be made will depend on the interest shown in the program as well as the contributions made by other researchers to this open-source software.

The most obviously desirable improvement is a refinement of the scoring system, to make it more linguistically informed. The best way to achieve this would be, in all likelihood, to use a parsed corpus, such that structures that typically intervene without disrupting a pattern, such as ellipses and relative clauses, could be treated differently from other types of intervening material. However, this would reduce the number of corpora with which SVEP could be used, and therefore we are somewhat reluctant to build in this requirement. However, a sophisticated analysis of POS data could cover much of the same territory, without placing any further requirements on the nature of the corpus.

The problem of requiring knowledge of regular expressions is somewhat trickier. Generally, there is a tradeoff between the flexibility of a search tool and the user-friendliness of its interface. Similarly, making things simpler for the user can lead to a tool being bound to a single corpus. While SVEP has thus far been used only with the BNC, it is designed to work – with minor modifications – with virtually any corpus, in any language. We wish to retain the flexibility and versatility that SVEP currently exhibits, at the same time that we would like to lower the threshold for using the program for different users. Currently, SVEP is distributed with sample files, including a sample regular expression file that defines certain useful sequences of regular expression material (similar to example 4 in Section 3.3 above). What could be done easily enough is to distribute the program with a larger set of codes and matching regular expressions, such that the user could use this list as a library and simply select the desired material for the investigation at hand. In this way, different libraries could be supplied for different corpora, making the process of composing regular expressions into a largely cut-and-paste exercise. This is an area in which different users of the program could contribute very helpfully.

A third improvement to the program, which would improve its flexibility, especially for the study of discourse-level phenomena, would be to relax the focus on the s-unit as the scope of searches. It might prove helpful to give the user the choice of matching patterns against s-units, paragraphs, or a certain number of words on either side of the node (a number which could be specified by the user). There is nothing to hinder the development of the program in this direction, if sufficient interest were expressed in this type of flexibility.

As mentioned above, it would certainly be an improvement to the program to add a graphical user interface, which most users have come to expect in software. However, this is likely to be added only if a great deal of interest is expressed, as it would require a considerable amount of development effort, and the program is offered as free software. Developers interested in designing a GUI would, of course, be welcome to contribute to the project.

5.3 Advantages and disadvantages of semi-automated approaches

The program SVEP has been discussed here in detail because it exemplifies the general approach to working with corpus data for which we argue here, termed "shared evaluation". We believe that this approach offers several distinct advantages. Ultimately, however, shared evaluation is only one example of what we would call semi-automated approaches to linguistic investigations.

We believe that the ideal approach to working with linguistic data is one that takes advantage of both the particular strengths of computers and the particular strengths of humans. The primary strength of a computer is that it is capable of performing a complex task many times at great speed. The primary strength of a human is the ability to make extremely complex analytical decisions, not least when it comes to making judgments about linguistic data. For this reason, in fairly large-scale studies, we prefer to do as much of the data extraction as possible using automated methods but then include at least one pass of "manual" evaluation of the data. If this manual evaluation can be speeded up by the further use of computers (see Garretson and O'Connor 2007), so much the better – but the main point is that the researcher him- or herself will eventually decide which tokens are in fact relevant to the study and how they should best be coded. Semi-automated approaches enable a research team to analyse as much data as possible, as accurately as possible, as quickly as possible.

Perhaps the most significant advantage of this division of labor between human and machine is that it leads to very high accuracy, with good precision and good recall (see Section 4.3). The final level of precision is high because the computer performs an initial classification, which sets the baseline for the manual phase much higher, making it easier to correctly classify the data. The final level of recall is high because the computer can be instructed to cast a wide net, with any irrelevant tokens being discarded in the manual phase. This allows a much higher level of accountability for the data in the corpus.

Another advantage is that semi-automated approaches make it possible to study much larger data sets than would be feasible otherwise. This, in turn, increases the confidence with which generalizations may be made, resulting in greater scientific rigor.

One disadvantage of semi-automated approaches is that they tend to involve a great deal of manual work with the data, especially if large data sets are used. This manual work can be facilitated, though, by aspects of the automated stage, such as the scoring system described above, which not only means that many or most tokens are already correctly classified but also points out those tokens that are most likely to be problematic.

Another disadvantage is the fact that implementing semi-automated approaches requires a number of technical skills. This means, in practice, that the research team will need to include someone with some programming skills. However, this need for technical skills can be alleviated somewhat if existing software is used as the point of departure.

Finally, it is worth considering that time put into the development of automated processes is likely to pay dividends in that these processes may be reused (for example, when the time comes to study a second corpus) or adapted to other uses (for example, to study a similar phenomenon) (see Garretson and O'Connor 2007). The program SVEP is a good example of this principle – although it was designed for a particular purpose, we are offering it to other researchers with the hope that it will prove just as useful for other studies.

Notes

[1] BNC*web* is available via a licensing agreement and typically runs on a server set up by the university obtaining this license.

[2] POS (part-of-speech) tags are tags applied to each word in a corpus by a program known as a POS tagger, which uses complex algorithms to determine as accurately as possible the syntactic class of each word.

[3] All the examples in this article followed by a three letter code (indicating the file name) are taken from the BNC. Examples with a dagger (†) instead of a file name are constructed or have been manipulated.

[4] Another fact about parsed corpora that is worth considering is that even the very best parsers still feature a considerable error rate. This means that any search that presupposes a correct parse of the sentence in question is likely to miss a sizeable proportion of tokens, which may be considered unacceptable for certain purposes. The program SVEP is designed (in the usual case) to make use of POS tags but not parse trees, as POS taggers typically feature a lower (though certainly still non-zero) error rate.

[5] *Svep* also happens to be a Swedish noun, meaning “sweep” or “swoop”, which seems appropriate, as the program sweeps through each corpus file several times, looking for different patterns.

[6] Strictly speaking, what is counted is the number of *tokens*, in the sense of *tokenization*, the splitting up of a corpus into word-like items; in many corpora, punctuation marks are separated from words, such that a string like *Well,* becomes two tokens. SVEP allows the user to decide whether both words and punctuation or only words should be counted.

[7] In the study reported here, we exclude cases of so-called ‘pseudo *tough*-movement’ (see Mair 1990: 72 ff.) as in “He is a tough man to please”.

[8] For those familiar with computer programming, it is helpful to think of these as the name of a variable and the value assigned to it.

[9] Use of spaces in the input files is not significant; all sequences of spaces or tabs in a pattern are considered equivalent, which means that spacing can be used to good effect to clarify the relationships between the patterns.

[10] The stop word list (see Section 3.4), which also functions as a filter, is input via a separate file.

[11] The errors made by the POS tagger were quite understandable; for example, it tagged *to police* not as an infinitival verb but as preposition + noun.

References

Anthony, L. (2011), *AntConc*. Tokyo: Waseda University. Available online at http://www.antlab.sci.waseda.ac.jp (last accessed on April 7, 2014).

Barlow, M. (2000), *Concordancing with MonoConc Pro 2.0*. Houston: Athelstan.

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

Burnard, L. (2007), *Reference Guide to the British National Corpus (XML-edition)*. Available online at http://www.natcorp.ox.ac.uk/docs/URG (last accessed on April 7, 2014).

Davies, M. (2004), *BYU-BNC* (based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc (last accessed on April 7, 2014).

Garretson, G. (2008), 'Desiderata for linguistic software design', *International Journal of English Studies*, 8: 67-94.

Garretson, G. and M. C. O'Connor (2007), 'Between the humanist and the modernist: semi-automated analysis of linguistic corpora', in: E. Fitzpatrick (ed.) *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse.* Amsterdam: Rodopi. 87-106.

Garside, R. and N. Smith (1997), 'A hybrid grammatical tagger: CLAWS4', in: R. Garside, G. Leech and A. McEnery (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman: London. 102-121.

Hoffmann, S. and S. Evert (2006), 'BNCweb (CQP-edition): the marriage of two corpus tools', in: S. Braun, K. Kohn and J. Mukherjee (eds) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang. 177-195.

Hoffmann, S., S. Evert, N. Smith, D. Lee and Y. Berglund Prytz (2008), *Corpus Linguistics with* BNCweb *– A Practical Guide*. Frankfurt am Main: Peter Lang.

Kaatari, H. (2012), 'Sampling the BNC – creating a randomly sampled subcorpus for comparing multiple genres'. Poster presented at *ICAME 33*, Leuven, Belgium, May 30 – June 3, 2012.

Mair, C. (1990), *Infinitival Complement Clauses: A Study of Syntax in Discourse*. Cambridge: CUP.

Mindt, I. (2011), *Adjective Complementation: An Empirical Analysis of Adjectives followed by* That-*clauses*. Amsterdam: Benjamins.

Nelson, G., S. Wallis and B. Aarts (2002), *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: Benjamins.

Reppen, R. (2001), 'Review of MonoConc Pro and WordSmith Tools', *Language Learning & Technology*, 5: 32-36.

Scott, M. (2012), *WordSmith Tools*. Liverpool: Lexical Analysis Software.

Van linden, A. (2012), *Modal Adjectives: English Deontic and Evaluative Constructions in Synchrony and Diachrony*. Berlin: De Gruyter Mouton.

Weichmann, D. and S. Fuhs (2006), 'Concordancing software', *Corpus Linguistics and Linguistic Theory*, 2: 107-127.

Using currency annotated part of speech tag profiles for the study of linguistic variation – a data exploration of the *International Corpus of English*

Marco Schilk

University of Hildesheim

Abstract

The International Corpus of English *has been widely used for the description of regional linguistic variation for the past two decades. The balanced corpus design of the ICE, which includes a large number of spoken texts and a large variety of different text-types and genres, however, also seems to be an ideal basis for the description of text-type differences in World English. In contrast to the tradition of solely focusing on variety-specific trends, this paper proposes using ICE as a basis to map out similarities and differences between regional varieties and different text-types. Data-driven in nature, it provides an exploration of nine different CLAWS7-tagged ICE subcorpora. After creating currency annotated part-of-speech tag profiles of the different subcorpora and the text-types and genres included therein, these profiles are used to identify homogeneous text-type groups. A comparison of the different groups makes it possible to isolate typical features of specific text-types but also points to some problematic issues concerning the design of the ICE.*

1. Introduction

During the last two decades the *International Corpus of English* (ICE) has been at the heart of many studies of variation between World Englishes (e.g. Hundt and Gut 2013). Early work on this project started in the 1990s and by now many different varieties of English have been included. These varieties can be roughly separated into native varieties of English spoken in monolingual or English-dominant countries (e.g. Great Britain, USA, New Zealand) and second-language varieties spoken in postcolonial and often multilingual communities (e.g. India, Hong Kong). Apart from the diversity of these varieties of English, the inclusion of many different text-types and genres and a well-balanced design that also contains a large proportion of spoken language data have made this corpus a valuable asset for everyone interested in the study of linguistic variation.

However, the goal of creating representative subcorpora of the varieties of English in question by including texts from the spoken and the written medium as well as many different text-types from various domains also entails several problems when focusing on regional linguistic variation. Apart from the fact that it is virtually impossible to create a truly representative corpus of any variety of English in general (cf. Leech 2006), doing so for a larger number of varieties creates additional problems to both corpus compilers and corpus users. Until today it has, for example, remained largely unclear, why corpora representing different varieties of English should contain the same proportions of different texts in order to represent the varying textual universes of the different varieties.

These global questions of corpus design are not easily addressed, since a concise estimation of any "textual universe" or total population the corpus is deemed to be representative of, is hardly possible, especially for a corpus that aims at containing samples of highly diverse text populations. However, there are a number of questions concerning the design of the ICE that can be put in focus more easily.

The first of these questions deals with the fact that the ICE contains several artificial 'dividing lines' along the dimensions of variety and text type. While it makes sense intuitively that it should be useful to investigate regional variation across a number of different text-types, the corpus design of the *International Corpus of English* seems to suffer from the fact that only a very small amount of text is included for each given text type. This may be one of the reasons why studies interested in register variation in varieties of English (such as e.g. Belasubramanian 2009) make only limited use of the ICE.

Researchers interested in regional variation, thus, face the trade-off of either using the complete ICE-subcorpora (or larger subsets such as the written and the spoken part of each component) regardless of register in order to have a relatively large sample size, or to compare very small register specific samples. However, it has often been pointed out that differences in register outweigh regional differences by far (Biber 1988: passim), so that collapsing data containing several different registers in order to arrive at a large sample is problematic, while at the same time the comparison of small samples, which in the case of ICE data often contain less than 20,000 words may not be useful to answer many linguistic questions. The present paper, therefore aims at shedding some light on the question of the importance of register variation in relation to regional variation with regard to the ICE. Moreover, the suggested method of using empirically defined text-type clusters, allows for collapsing different similar text-types into larger datasets in those cases where the ICE-subcorpora do not contain sufficient data within the specific text-type categories.

A further question addressed in this paper deals with the relation of spoken language and written language represented in the ICE. Keeping in mind that one of the greatest benefits of the *International Corpus of English* is the relatively large amount of spoken language that is part of each subcorpus, the corpus dividing line between spoken and written text categories is also put under scrutiny. More specifically, the question of medial versus conceptual spokenness/writtenness will be put in focus, since the underlying corpus design of the ICE focuses rather strongly on medial differences and to some degree disregards conceptual distinctions. It is, however, likely that some of the included written text types (such as e.g. creative writing) more closely resemble medially spoken language, while other text types that are sorted under spoken language closely resemble conceptually written language, most likely in the scripted spoken text types, such as broadcast news reportage.

The present paper uses a quantitative method based on currency-annotated part-of-speech tag profiles in order to show which texts in the different subcorpora of the ICE show statistically significant similarities to other texts. This methodology makes it possible to show in how far register variation outweighs regional variation and provides information concerning the questions on medial versus conceptual spokenness/writtenness raised above.

2. Database and methodology

The dataset underlying the present pilot study contains four native varieties of English (ENL) (British English, Irish English, Canadian English and New Zealand English)[1] and five institutionalized second language varieties of English (ESL) (Indian English, Singapore English, Hong Kong English, Philippine English and Jamaican English). This selection is mainly based on practical considerations. Only those corpora were included that exist in a finished form at the time of writing this article and contain both written and spoken data, as one objective of the present paper is to investigate spoken/written distinctions in terms of medial versus conceptual spokenness/writtenness. This led to the fact that some major varieties, such as American English are not included in this pilot study.

Concerning the second-language varieties, the data has a certain bias towards Asian Englishes. This is due to the fact that it is the Asian corpora that are the most complete second language subcorpora in the ICE project. Exceptions are ICE-Jamaica and ICE East-Africa. ICE East-Africa, however, is not included in the present analysis, as it differs considerably from the other ICE-subcorpora in terms of the underlying policies of text selection, so that this ICE corpus is not directly comparable to most other ICE-subcorpora. Table 1 gives an overview of the included corpora.

Table 1. ICE-subcorpora

Corpus	Size	Status of English
ICE-Canada	all corpora contain approx. 1 million words 600,000 spoken 400,000 written	ENL
ICE-Great Britain		ENL
ICE-Hong Kong		ESL
ICE-India		ESL
ICE-Ireland		ENL
ICE-Jamaica		ESL
ICE-New Zealand		ENL
ICE-Philippines		ESL
ICE-Singapore		ESL

For the creation of POS-tag profiles all corpora were tagged with the CLAWS7 tagset (Garside 1987).[2] Moreover, all lexical items contained in the corpora were annotated for lexical currency. In this study, lexical currency is an index of membership of a lexical item in the number of different corpora used in the study. Thus, central items are items that are used in the majority of the corpora under scrutiny (the *lexical core*, in other words) whereas peripheral items are those only used in a small number of corpora. This procedure makes it possible to identify the proportions of central and peripheral items used in each corpus. For example, most closed-class elements are relatively central in the sense that they occur very frequently in all of the corpora. Less central are open-class elements, such as nouns and lexical verbs that occur less frequently and display a lower dispersion across corpora. On the peripheral end we find those items that are either used only in a single corpus due to regional variation (i.e. the typical X-isms) or those items that are used so infrequently in general that they only occur in one of the corpora. Examples of these items are rare lexical nouns and verbs, such as the verb *quench* or archaic and obsolete close-class items such as e.g. the determiner *yon.*

While the initial currency-check is based on word-type lists, the currency index is added to all tokens in the specific corpus. The annotated items are further sorted along the lines of closed-class and open-class lexical items, since it is likely that most closed-class items are very current and, in the majority of cases, quite frequent, whereas lexical innovations at the periphery will probably be largely restricted to open-class items.

This annotation of lexical items makes it possible to create lexical profiles for each of the subcorpora and for each of the text-type categories included in each subcorpus, as visualized in Figure 1).[3]

Figure 1 is a representation of the process of identifying lexical profiles for texts and also collections of texts, i.e. corpora. The oval on top represents any given text that is subject to profiling. This can be a relatively small single text, such as, for example, a single personal letter, or a large collection of texts, as with a complete electronic corpus. Depending on the type of text profiled, there will be varying degrees of granularity of the profile. If a single short text is being profiled, naturally the lexical profile for this text will be highly specific, while profiles for large text collections are more representative but will contain less resolution with regard to any of the specific texts included.

Each text is composed of items coming from nine different lexical pools, i.e. lexical items varying in degree of currency. The different pools indicate in how many of the underlying ICE-subcorpora the item is used, so that membership in any of these pools is an index of lexical currency.

Each of these pools is additionally subdivided into closed-class and open-class items. Thus, any given text has a specific profile depending on which parts of speech are used, how common the specific words in the text are, and the relation of open- and closed-class items, which can be seen as a way of measuring lexical density of a text. The size of the arrows (although not drawn to scale) indicate the fact that most texts contain a relatively high number of very current lexical items, and considerably fewer items that are restricted to a smaller number of source corpora.

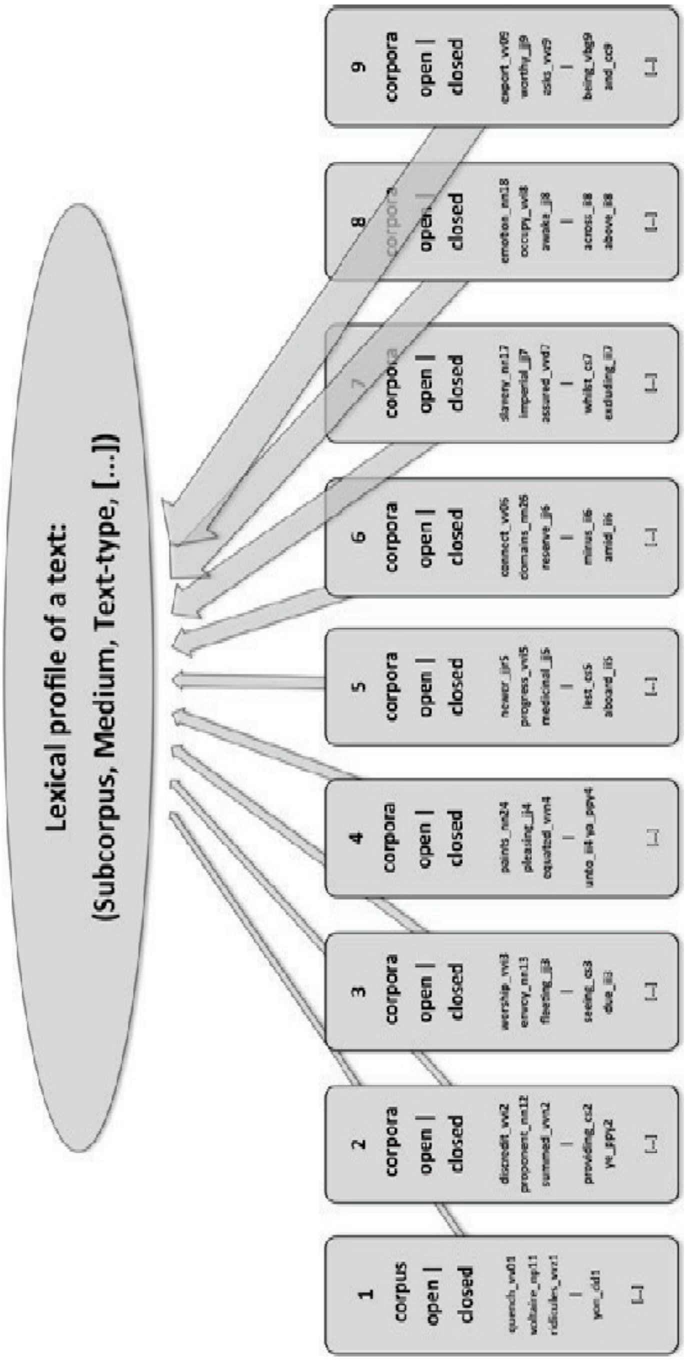


Figure 1. Currency-based lexical profiles

For example, *quench*, in the left-hand column, is a relatively rare lexical verb that only occurs in one of the ICE-subcorpora. Of course this does not suggest that *quench* is uniquely used in one variety of English. On the contrary, with a corpus size of 1 million this item is so rare that it only occurs in a single corpus (and most likely only in collocation with *thirst* or a metaphorical extension thereof). The same holds true for the use of the closed-class determiner *yon*. This item is also highly restricted in use and therefore only encountered in one of the corpora.

At the other extreme (rightmost column) are those items that are found in all corpora. These are basically words with a high functional use and versatility. Examples here are the lexical verb *ask* that is used with virtually all formulations of a question and the conjunction *and*, the use of which is associated only with certain grammatical but no lexical restrictions.

Based on these profiles it is possible to compare the lexical composition of different texts and identify similarities and differences statistically. However, since these comparisons are based on distance matrices, using the complete entries for each word is not feasible. This is due to the fact that the computational power needed to calculate the respective matrices is immense. A solution to this problem is the use of currency-annotated POS-tag profiles. These profiles are basically identical with the complete lexical profiles with the exception that the actual lexical items are no longer represented. Thus, for example, the rare lexical verb *quench*, that is represented in the lexical profile as *quench_vv0-1* (*quench* as a lexical verb (vv) in its base form (0) that occurs in one corpus (1)) is represented as _vv0-1 (a lexical verb (vv) in its base form (0) occurring in only one corpus (1)) in the POS-tag profile). Note that this procedure allows for a more fine-grained picture than a non-currency-annotated POS-tag profile, since it includes some of the properties of the item that are not part of the part-of-speech profile (e.g. the rareness and, by extension also the restricted versatility of the verb *quench).*

The corpus analysis in Section 3 is carried out in three parts. Firstly, an overview of the different lexical composition of each corpus with regard to the currency of the lexical items is presented. In a second step, the larger subcorpora are split up according to the different text-type categories predefined in the ICE. As these text-categories are defined on functional rather than on empirical linguistic grounds, in a next step, a cluster analysis based on the currency-indexed POS-tag profiles is performed to illustrate how different functionally defined text-type categories display similarities in their lexical composition.

This cluster analysis is based on the following assumptions.

(1) The proportional preference for the use of specific lexical items and, by extension, the proportion of currency-indexed parts-of-speech differ in different sorts of texts (e.g. according to medium, text-type, genre or register). This has been pointed out in earlier studies such as for example in the work of Biber (e.g. Biber 1988, Conrad and Biber 2001).

(2) By analysing the relation of currency-indexed POS-profiles it is possible to build empirically defined groups of texts. These groups may cross traditional dividing lines, for example, between the written and the spoken medium. This is in line with the notion that some medially written texts are conceptually closer to the spoken medium and vice versa. Furthermore, similarities between functionally different genres can be identified on empirical grounds.

(3) Since texts within a cluster are significantly similar with regard to their lexical/POS composition, the text-types within a cluster are defined as an empirically derived new text-type. Note that this new text-type is only based on the similarities in lexical/POS composition by collapsing the texts within a cluster under the new text-type header.

(4) Variation between text-types can be identified by studying the quantitative differences in the composition of different types of words represented by their respective currency annotated POS-tags. By focusing on those items that differ most markedly between different texts, it may be possible to explain the different character of text-types.

For the present analysis these assumptions are put to the test on the basis of the ICE-subcorpora described above. The cluster analysis is performed with the R package {pvclust}. The hierarchical clustering is based on Ward's sum of squares method (Ward 1963), multiscale bootstrap resampling with 1000 replication is used to test for statistical significance (approximately unbiased probability values (AU)) (Suzuki and Shimidora 2006).[4]

3. Results

The cluster analysis in the current study abstracts from the complete currency-annotated lexical profile and uses a truncated version of this profile that does not contain the actual lexical items represented in the corpora. This truncated version still represents each lexical item used, but only differentiates between part-of-speech and lexical currency. Thus, if the item *quench_vv01* occurs eight times in a given text, these occurrences are represented as eight uses of a peripheral lexical verb occurring in its base form (*vv01).* The currency-annotated composition of each text, therefore, allows for the statistical identification of similarities and differences between different texts and text-types.

In order to find out if different texts display similar lexical composition, the corpora are split up into their component text categories. Table 2 gives an overview of the text-types included in the ICE with their respective textcodes:

Table 2. The *International Corpus of English* (Nelson et al. 2002: 307-308)

Spoken (300)	*Dialogues (180)*	Private (**S1A**) (100)	Conversations (90) Phonecalls (10)
S1A/S1B		Public (**S1B**) (80)	Class Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Cross-examinations (10) Business Transactions (10)
S2A/S2B	*Monologues (120)*	Unscripted (**S2A**) (70)	Commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		Scripted (**S2B**) (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)
Written (200)	*Non-printed (50)*	Student Writing (**W1A**) (20)	Student Essays (10) Exam Scripts (10)
W1A/W1B		Letters (**W1B**) (30)	Social Letters (15) Business Letters (15)
W2A-W2F	*Printed (150)*	Academic (**W2A**) (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Popular (**W2B**) (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Reportage (**W2C**) (20)	Press reports (20)
		Instructional (**W2D**) (20)	Administrative Writing (10) Skills/hobbies (10)
		Persuasive (**W2E**) (10)	Editorials (10)
		Creative (**W2F**) (20)	Novels (20)

As can be inferred from Table 2, the text-type organization of the ICE-corpora can be viewed at different levels of granularity. On the most macroscopic level there is a distinction between spoken and written data included in the corpora. Each corpus contains 600,000 words of spoken English and 400,000 words of written English. However, it should be kept in mind that this distinction is purely medial. Thus, whenever sound files have been recorded and transcribed, the respective texts are included in the spoken category; whenever texts have been available in written format, they are sorted under the written category. This has some implications for the types of texts included in the corpus, as written-to-be-spoken texts (like e.g. all scripted monologues in the S2B category) are, despite being medially spoken, conceptually written texts.

This brings us to the next level of granularity, i.e. the subdivision into dialogues and monologues in the spoken data and the corresponding subdivision into printed and non-printed material in the written section. Keeping in mind the conceptual/medial distinction, it could be argued that dialogue data are more prototypical of conceptually oral language and that printed written data is more prototypical of conceptually written texts than their respective counterparts, monologues and non-printed texts.

The next smaller level is situated at the interface of text-type and genre and the category titles are not entirely homogeneous or stringent. While at the level of dialogue data the main distinction is participant/audience oriented (private vs. public dialogues), at the level of monologue data the process of text production is in focus (scripted vs. non-scripted). In the written part of the corpus the distinctions are closer to genre distinctions, covering categories such as letter writing, academic writing, reportage or creative writing.

This level is the main coding level for the corpus texts, meaning that each text has a code assigned to it that makes it directly identifiable as belonging to a certain category. This level is also the level chosen for the cluster analysis below, as it provides a relatively high resolution but does not include so many subgenres as to make the results too unwieldy to interpret. At the level that contains the highest differentiation the corpora are further divided along the lines of different genres and different topics (e.g. in the academic writing category different academic disciplines are represented).

Since one of the motivations driving the present analysis is to provide a bottom-up approach to the distinction of media and text-types in different varieties of English, the analysis starts by dividing the texts at the second-highest level of resolution, namely S1A to W2F.

Figure 2 shows the cluster analysis of currency-annotated POS-tag profiles of the different open-class items in the text-type categories S1A to W2F in the different variety-based subcorpora of ICE. This dendrogram contains several noteworthy points that will be discussed in some detail in the remainder of this paper.

Firstly, Figure 2 shows that variety does not seem to be a highly distinguishing factor. In most of the clusters many different varieties of English are represented and there also does not really seem to be a significant dividing line between native and second-language varieties of English. There is, however, one exception to this general observation. As can be seen on the right side of the dendrogram in Figure 2, the spoken New Zealand data forms a cluster of its own. This runs somewhat contra to the assumption that text-type differences are more influential for variation than regional differences. Thus, this data and the possible differences between spoken NZE and the other varieties will be looked at more closely below.

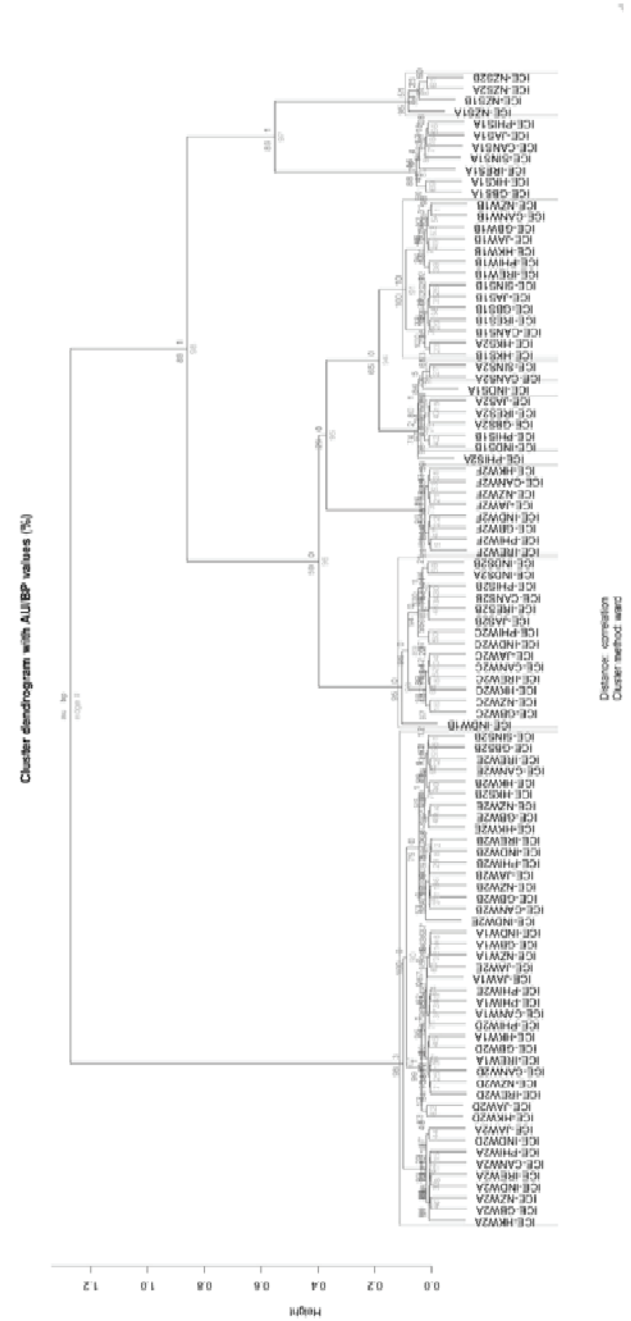


Figure 2. Cluster analysis of open-class currency-annotated POS-tag profiles by variety and text-type[5]

Secondly, the dendrogram shows a rough distinction between text-types that are more characteristic of written language and those that are more characteristic of spoken language. While this is not particularly surprising, given the fact that the main distinction between the different ICE categories is the distinction between spoken and written texts, a closer look at the graph shows that not all texts on the left-hand side of the dendrogram are medially written texts (coded W) and not all texts on the right-hand side are medially spoken texts. Therefore, for some of the texts a conceptual interpretation of the written vs. spoken distinction may be more adequate than a medium-based distinction.

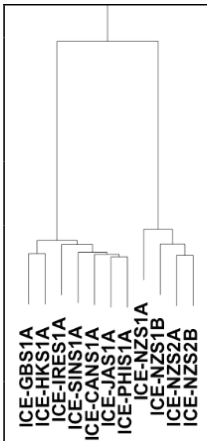
This aspect figures most prominently in the second significant cluster from the right, which contains unscripted spoken monologues and letter writing (S1B and W1B), and in the second significant cluster from the left, which contains both written and spoken news data.

Consequently, the following in-depth analysis of the different POS-tag profiles concentrates on the three clusters mentioned above. Firstly, the differences between the spoken New Zealand data and the spoken dialogue data of the other varieties (as the 'closest neighbour') are focussed on in the analysis. For the sake of identifying the differences and similarities of those texts, the four texts of the NZE data are compared to the seven S1A texts that resemble those texts most closely but do not form a significant cluster with the NZE data. In the analysis both branches of the tree, i.e. the four NZE texts and the seven S1A texts are treated as single texts and the differences between them are compared statistically in order to identify those parts-of-speech that have the strongest influence on the difference between the currency-annotated POS-tag profiles. Table 3 gives an overview of the major differences between the NZE spoken texts and the control texts.

The information in Table 3 can be read in the following way. It contains those items that display the largest variation between the two branches of the dendrogram shown on the left. The column labelled "tag*"* provides the information of the CLAWS7 POS-tag and the currency annotation. The column labelled "type" is the translation of the currency tag information, e.g. *nn11* in the tag-column stands for a singular common noun (*nn1*) that is only found in one of the corpora (*1*).

The numerical values are the residuals of the comparison of the two datasets. Positive or negative values indicate overrepresentation or underrepresentation with respect to the other dataset. In other words *nn11* items are more frequent in the spoken NZE data than in the rest of the spoken dialogues. The column *percentage of variation* uses the sum of all squared residuals and displays the percentage of the squared residuals and, thus, the proportion of the variation for each row. This means that, since the total variation between all elements in the distribution has a value of 73,426.37 (sum of squared residuals), the sum of squared residuals in the first row (*nn11*), 13,803.94, equals 18.8% of the total value. Therefore, the ten items represented in Table 3 are responsible for about 74% of the differences in usage of lexical items between the two datasets.

Table 3. Comparison of spoken dialogues (S1A)

	Tag	Type	NZS	S1A	Var. (%)
	nn11	singular common noun / 1 corpus	94.4	-69.9	18.80
	mc1	cardinal number / 1 corpus	78.5	-58.2	13.02
	fu6	unclassified word / 6 corpora	72.1	-53.4	10.97
	np11	singular proper noun / 1 corpus	-68.6	50.8	9.92
	vv09	base form of lexical verb / 9 corpora	-55.7	41.2	6.53
	np18	singular proper noun / 8 corpora	-41.5	30.7	3.63
	mc19	singular cardinal number/ 9 corpora	41.3	-30.6	3.60
	nn18	singular common noun / 8 corpora	-37.0	27.4	2.89
	vvi9	infinitive / 9 corpora	-33.5	24.8	2.36
	vvz7	s form of lexical verb / 7 corpora	30.5	-22.6	1.96
	SUM				73.67

Hence, by interpreting the ten different lines of Table 3, it is possible to explain the main differences between the NZE spoken dialogue data and the S1A data of the other corpora.[6] For the first line, which shows the difference in the distribution of peripheral common nouns, a look at the source data shows that there is a strong difference in the transcription conventions applied to the spoken data in ICE-NZ and the other corpora. While in the other corpora the conventions for orthographic transcription follow the rules of written English rather closely, it seems that in the NZ data the difference of spoken and written language is reflected more strongly in the transcription conventions. One consequence of the difference in annotation is that there is no capitalization in the spoken part of the corpus. This produces the result that certain items, like the days of the week, are sometimes classified as common nouns by the CLAWS tagger, while they are classified as proper nouns in the other corpora. Since these falsely identified common nouns do not exist (as common nouns) in the other corpora these items are annotated as peripheral items, i.e. items that can only be found in one corpus.

The next difference between the two datasets, peripheral singular proper nouns (np11), is rooted in the comparison of a single ICE corpus with a set of different corpora, since it is quite natural that you will find more corpus specific names of people, places, brands and so on in the collapsed data of seven corpora compared to one corpus. In other words, the NZ dataset only contains those peripheral proper nouns that are used in ICE-NZ, while the other dataset contains rare proper nouns of seven different corpora.

The higher frequency of use of np18 and nn18 is, again, due to differences in corpus transcription and resulting differences in annotation. Hesitation markers in ICE-NZ are usually annotated as interjections, while CLAWS seems to interpret the hesitation markers in the other spoken corpora as common or proper nouns, depending on their position in the sentence. Thus, uhm is either interpreted as a common noun or a proper noun in eight of the corpora, although not in ICE-NZ, which leads to a higher frequency of np18 and nc18 in the pooled data.

Finally, there seems to be a more frequent use of the s-form of lexical verbs that occur in seven corpora in ICE-NZ. Again this has to do with differences in corpus transcription, as in ICE-NZ laughter is marked as laughs, a verb-form that does occur in this form in seven corpora. The use of this item for corpus annotation in ICE-NZ, thus, leads to the higher frequency of usage of vvz7 forms in ICE-NZ.

Thus, almost 54% of the difference between the two datasets is a based on transcription differences, incompatibility of these differences with the tagger that has been used, the annotation algorithm, or unclassified words and numerals. Therefore, it can be assumed that without these, the spoken dialogue data in NZ will not differ significantly from the same data in the other corpora. As a consequence, in the only case where "variety" seemed to have a significant impact on clustering, the differences were based on different corpus transcription rather than language-inherent features. This strengthens the assumption that different varieties are relatively homogeneous within different text-type categories based on the lexical profiles of the different texts. It also illustrates the importance of adherence to comparable transcription guidelines when compiling a corpus that is part of a larger corpus family.

The second sample analysis within the present pilot study deals with the categories that include news reportage (S2B and W2C). There are several noteworthy points concerning the texts of these categories that are grouped together in a significant cluster. The main observation concerning these texts is the fact that they are very similar in terms of lexical composition, although they belong to medially different text-types. While this is not particularly surprising, considering the nature of both spoken and written news reportage, the bottom-up grouping of these two text-types by the cluster analysis serves to illustrate at least two points.

Firstly, the unsurprising nature of this observation highlights the strength of the bottom-up approach used in this study. Despite the limitations of the underlying methodology that were pointed out in the previous discussion of the NZ data, in general the use of currency-annotated profiles for the identification of text-type similarities seems to perform rather well, if text-types that are intuitively very similar are grouped together. Furthermore, this group also serves to corroborate two of the hypotheses formulated earlier. The first of these hypotheses was that conceptual properties of texts seem to outweigh the medium when it comes to the differentiation of written versus spoken data. Although the texts within this group are medially spoken in the case of the S2B texts and medially written in the case of the W2C texts, the informational nature of both texts is highly similar, which is reflected by a similar distribution of currency indexed parts of speech.

The second hypothesis states that usually differences between the respective text-types are far more pronounced than regional variation, so that grouping is more likely to occur at the level of text-type than at the level of variety. The dendrogram in Figure 5 largely corroborates this hypothesis and especially the news text categories (and the fictional texts in the W2F) serve as cases in point. However, there are some exceptions to be considered, because some news categories are not found in this cluster (S2B in ICE-GB and ICE-SIN) and unscripted Indian monologues are also part of this cluster.

In order to investigate why the Indian monologues are grouped among the news texts rather than together with other unscripted data, it is useful to compare these texts with other unscripted data rather than with other news texts. The reason for this is that the unscripted Indian monologues in question are significantly similar to the news texts in the other corpora, so that a comparison of those text-types will not yield significant differences. On the other hand, a comparison with other unscripted monologues shows how these texts are different from those and why they may be grouped with the news text. This analysis is carried out in the same fashion as the comparison of the spoken dialogue datasets above. In this manner, the profile of the Indian unscripted monologues is compared to the profile of British unscripted monologues. The decision to use only one variety for this comparison is based on the difficulty in comparing a single corpus to a number of corpora, which became apparent during the analysis of the NZ spoken data (e.g. in the case of np11 items).

Table 4. Comparison of unscripted monologues (S2A) in ICE-India and ICE-GB

Tag	Type	ICE-GB S2A	ICE-Ind S2A	Var. (%)
np11	proper noun sg. / 1corpus	-21.65	20.22	24.46
np16	proper noun sg. / 6 corpora	12.36	-11.54	7.97
nn11	common noun sg. /1 corpus	-8.60	8.03	3.86
np12	proper noun sg. / 2 corpora	8.32	-7.77	3.62
np13	proper noun sg. / 3 corpora	8.32	-7.77	3.62
np15	proper noun sg. / 5 corpora	8.00	-7.47	3.34
np17	proper noun sg. / 7 corpora	6.92	-6.46	2.50
SUM				49.36

Table 4 shows the main differences between the unscripted monologues of ICE-GB and ICE-India. A closer look at this table gives a good indication of the reasons for the grouping of the Indian texts together with news reportage. As the table shows, the main difference between Indian unscripted monologues (since they are significantly similar to the news texts) and British unscripted monologues lies in the use of nouns. Especially peripheral proper nouns are used much more frequently in the Indian unscripted monologues compared to the British reference texts. The reason for this seems to lie in the inclusion of a large amount of cricket commentary in the Indian corpus. These commentaries contain a large amount of specifically Indian names of people and places. Since the use of names of people and places particular to a specific region are also typical of news reportage, this high proportion of proper nouns in ICE-India is the most important reason for those texts to appear “newsier” than the British texts in the same S2A category. Although the British S2A texts also contain a significant amount of sports reportage, many of the British names for people and places are not as peripheral as those in the Indian texts. Thus, typical British names will also be found in sports reportage in the other corpora (as the overrepresentation of more central proper nouns in the British data shows), while the peripheral Indian names in the Indian corpus are closer to the region specific usage in news reportage in the other corpora.

The third sample analysis in this pilot study differs from the previous two by adopting a slightly different perspective. In this analysis much less emphasis is put on regional variation but the differences between spoken and written data are put more strongly into focus. This analysis is based on the observation that public dialogues and letters are grouped together. As in the case of the news data, medially written and spoken data display significant similarities. Therefore, in a first step, we will look at those items that are most frequent in both of the categories (S1B and W1B). Because they form one significant group, it is to be expected that the most frequent items in both of the text-categories display considerable overlap and that we will get a first indication of which items are typical of this cluster.

In a second step the cluster that contains S1B and W1B texts will be treated as a single text-type and be compared to a text-type that is prototypical of texts that are both medially and conceptually written, namely academic written texts (W2A).

Table 5 displays most frequent items in the lexical profiles of the texts in the S1B and W1B category.

Apart from the obvious fact that the profiles of S1A and W2B are very similar, the profiles in Table 6 also give a good indication of the (content) items that are very typical of dialogue texts and letter writing; especially when it comes to the verbal forms, we can see a clear preference towards central infinitives and base forms. We can further identify a high number of central general adjectives, central common nouns and peripheral proper nouns. Concerning verbal usage, the frequency of base forms and infinitives indicate a preference for the present tense, the use of the past participle indicates use of the present perfect and the past perfect rather than the simple past. The high number of peripheral proper nouns is also typical of both dialogue and letter writing, since these texts are usually directed towards single (named) interlocutors.

Table 5. Lexical profiles of S1B and W2B (top ten item-classes)

S1A (N=344786)			W1B (N=204857)		
tag	type	#	tag	type	#
nn19	singular common noun	71364	nn19	singular common noun	41216
jj9	general adjective	28261	jj9	general adjective	16567
vvi8	infinitive	22338	vvi9	infinitive	12290
vv09	base form of lexical verb	20574	nn29	plural common noun	8980
nn29	plural common noun	16184	vv09	base form of lexical verb	7686
nn18	singular common noun	13047	np11	singular proper noun	5832
np11	singular proper noun	9224	nn18	singular common noun	5657
vvn9	past participle of lexical verb	9003	vvn9	past participle of lexical verb	5367
mc9	cardinal number	8218	vvg9	-ing participle of lexical verb	4558
vvg9	-ing participle of lexical verb	8118	nnt19	temporal noun, singular	4454

Table 6. Comparison of conceptually spoken texts (S1BW1B) and academic writing (W2A)

tag	type	S1BW1B	W2A	Var. (%)
vv09	base form of lexical verb	67.0	-81.1	14.3
vvi9	infinitive	59.2	-71.5	11.1
mc1	cardinal number	-44.2	53.4	6.2
rt9	quasi-nominal adverb of time	37.0	-44.7	4.3
rl9	locative adverb	30.0	-36.2	2.9
vvg9	-ing participle of lexical verb	28.9	-35.0	2.7
vvd9	past tense of lexical verb	28.4	-34.3	2.6
nnt19	temporal noun, singular	28.2	-34.1	2.5
rg9	degree adverb	28.1	-34.0	2.5
SUM				49.0

In the second analytical step, the data of letters and public dialogues is pooled and compared to a text category that can be seen as a prototypically written category, academic writing, in order to identify how conceptually spoken data differs from conceptually written data in terms of currency-based lexical profiles.

The comparison in Table 6 is basically in line with the observations drawn from the analysis of the lexical profiles of the conceptually spoken texts. The main feature that distinguishes the conceptually spoken texts from the prototypical written text is the use of lexical base and infinitive forms. On the one hand, this may be due to a tendency to use present tense forms more frequently in spoken data than in writing. However, we can also identify a tendency to use central lexical verbs in the past tense more frequently in the conceptually spoken data.

Thus, tense is not the only difference in verbal use between those two text-types. The other differences may well be grounded in the shorter sentence length overall in spoken language, which leads to a higher proportion of verbs. Furthermore, in the conceptually spoken text-types very central verbs are used more frequently compared to academic writing, where verbs are also more specialized and therefore more peripheral in terms of inter-corpus currency.

Another observation of this analysis is the fact that almost all differences are in categories that cover central items, i.e. items used in all corpora. The only exception here is the higher frequency of use of cardinal numbers that only occur in one corpus in the W2A data. This use is quite plausible, since reporting on scalar data in academic writing will make use of peripheral cardinal numbers.

Therefore, the analysis of S1A and W1B texts and the comparison of those texts as a pooled dataset for academic writing illustrate several points. Firstly, we can see that there is frequent use of central content forms in the conceptually spoken data. In fact, the use of forms that have a high inter-corpus currency seems to be one of the defining factors of conceptually spoken texts. Concerning medially spoken texts, this is not particularly surprising due to speakers' tendency to use frequent, short, more general and versatile forms in real-time speech situations. While more peripheral forms may carry a higher semantic weight in the sense of being less vague or ambiguous, language economic factors seem to outweigh the deficiencies of more general forms over highly specific, rare and non-versatile forms.

In the case of letter-writing, however, economic considerations relevant for real-time speech situations are far less important, so that the reasons for the high similarity of content-item use between the two text-types may need some further explanation. Especially in the case of private letter writing, the fact that writer and addressee usually know each other relatively well may facilitate the use of those more general forms, since the personal relationship may be a factor that eliminates possible vagueness, so that the correspondents do not have to use highly specific and unambiguous language.

Topic may be a further factor, since personal letters often refer to experiences of the writer in his or her daily life. Thus, experiences that can be seen as cognitive prototypes play a central role in the topics of letter writing, which may also serve to explain the similarities of those texts with spoken data, where the interlocutors directly interact and can repair possible misunderstanding in a more direct fashion.

The text-category academic writing, which was chosen as a prototypical written text-type, both medially and conceptually, is in stark contrast to both public dialogues and letter writing. In this category far fewer central verbs are being used. This may be motivated by several factors. Firstly, there is an assumed tendency to use more specialized semantically heavy verbs, as pointed out above. However, by looking at Table 6, it seems that at least on the level of the ten lexical categories that differ most strongly between the text-types, there is no comparative overuse of rare forms of peripheral verbs in the academic writing category. The reason for this may simply lie in the differences in sentence length and, therefore, the total amount of verbs being used (cf. e.g. Halliday 1989).

Since the number of verbs used in a sentence does not grow proportionally to the length of a sentence, texts consisting of short sentences will automatically contain a higher proportion of verbs than longer sentences. Although the underlying methodology of this paper does not take sentence length into account, the results in Table 6 clearly point in this direction, since those verb forms that are characteristic of the conceptually oral text-types are among the categories with the higher amount of variation, while the only comparative overrepresentation of a lexical category in academic writing is attested for the usage of peripheral cardinal numbers.

As the present methodology only compares proportional use of currency-annotated POS-tags, some of the differences between text-types, such as e.g. sentence length, and their influence on the different lexical/POS-tag profiles are not explicitly included. This is of minor relevance when creating text-type clusters, since similar text-types are also similar in the proportional use of items and similarities in sentence length, type-token ratio, etc. are accounted for. Nevertheless, when comparing very different text-types to each other, it is only possible to describe the numerical differences of use, while the reasons for these differences cannot really be quantified. In order to arrive at a more fine-grained analysis of the underlying reasons for the differences in profile composition, it might, therefore, be useful to discount effects of sentence length, type-token ratio and other influential factors, an approach that has not been used in the current pilot study.

4. Conclusion and outlook

The present paper has introduced a methodology to compare texts from different varieties and text-types in a bottom-up approach by creating currency-annotated POS-tag profiles. Based on nine different subcorpora of the *International Corpus of English*, this method was used to investigate several questions important for the use of the *International Corpus of English* in studies of regional and text-type variation as well as corpus design and creation from a wider perspective.

The corpora were separated along the lines of variety and text-type. By creating currency-indexed POS-tag profiles it was shown how specific text-types form significantly similar text-type clusters. The dividing lines between those clusters were mainly between conceptually written and conceptually spoken text-types; regional variation only played a minor role in the grouping of texts. In the case of NZE spoken data it seemed as if this regional variety differed significantly from the other spoken varieties. However, a closer look at the text-type profile showed that a large part of the variation between NZE dialogue data and the equivalent spoken data of the other regional subcorpora is mainly based on differences in corpus design and annotation.

Accordingly, the main concluding point of this analysis is the plea for more consistent corpus design in the different subcorpora. Since, differences in data collection and annotation between the different subcorpora of ICE are often not easily identifiable at first sight, a more stringent and diligent corpus documentation is clearly needed to make full use of the possibilities offered by the *International Corpus of English*. As matters stand at the moment, with corpus documentation kept at a bare minimum, researchers may report on differences in corpus design rather than reporting on variety-based differences without being aware of the fact, which is clearly undesirable.

Apart from the fact that regional variation seems to be relatively marginal compared to text-type variation on the level of POS-tag profiles, the second part of the analysis showed that the empirical dividing line between spoken and written texts differs from the artificial dividing line proposed by the ICE corpus design where the definition of spoken versus written text in the ICE corpus is purely medial. However, much of the medially spoken data in the corpus more closely resembles written data, while some written text types closely resemble conceptually spoken language. In order to make full use of the ICE-corpus it may, therefore, be useful to reorganize the corpus into text-type groups that more closely resemble the conceptual written/spoken distinctions that are suggested by the results of the current cluster analysis.

I want to close my conclusions with some further remarks on the possibilities and limitations inherent in the *International Corpus of English*. Apart from these issues concerning corpus design discussed above, the analysis has shown that the *International Corpus of English* is a very useful corpus collection with applications that go far beyond the mere description of peripheral variety-based variation. Especially the inclusion of a large spectrum of different text-types open up possibilities for studies that go beyond traditional variational linguistics. Since there is an immense overlap of core features of the different varieties, using the corpus family as a whole opens up possibilities of describing features of the English language without being subject to regional variation.

However, there are some questions that I feel need to be answered to conserve the merits of this corpus for the future. Firstly, the macrodesign of ICE is not based on any overarching linguistic considerations, so that the question which regional varieties are added to the corpus family is basically answered arbitrarily. Due to the resource-consuming nature of creating an ICE subcorpus, the ICE-community understandably welcomes any addition to the existing data, regardless of macrodesign. Thus, if interest arises in creating a subcorpus of a relatively esoteric variety, this is usually met by unanimous applause and considerations of a balanced corpus design with respect to the regions represented in the corpus are not taken into account.

While this 'grassroots' approach seems to have worked quite well in the past, it has also led to the situation that some major varieties of English are not completely represented and that some subcorpora have become relatively dated, creating an inner-corpus diachronic gap. A prime example of a major variety that is not yet included is American English. Although the written component of ICE-USA has recently been published, no publishing date for the completion of the spoken component exists at this stage and, thus, a completely comparable subcorpus of American English will probably not be available for a number of years. This situation is even more unfortunate, since the separate publication of a written and a spoken component is also almost guaranteed to result in a diachronic gap within ICE-USA, with the texts included in the written part being much older than those included in the spoken part.

For improved comparability of the ICE-subcorpora, it also seems to become increasingly necessary to update the data available in the older corpora. Although the present analysis has shown that the time-gap between the corpora does not seem to have an immediate influence on the composition of specific texts, text-type evolution (e.g. from letter writing to electronic communication) will become an increasingly problematic issue for corpus comparability. Thus, an inclusion of a complete ICE-USA and an update to ICE-GB should be a very high priority on the agenda of anybody considering contributing to the *International Corpus of English*.

Notes

[1] It should be noted that ICE-Can does not only represent ENL speakers. However, the amount of material contributed to ICE-Can by non-native speakers of English (with mainly French as their L1) is only about 10%. Thus, although Canada is a special case, it seems to make sense to sort Canadian English among native varieties of English.

[2] The corpus files have been automatically tagged horizontally with the *jclaws* tool of the Lancaster CLAWS tagger. No manual correction of the resultant tagged corpora was undertaken, as such a procedure was not feasible within the scope of the present pilot study for approx. 9 million words of data.

[3] Note that *hapax legomena* have been excluded from the analysis.

[4] Generally speaking, for Ward-type clustering, "[t]he clustering criterion is based on the error sum of squares, *E*, which is defined as the sum of the squared distances of individuals from the centre of gravity of the cluster to which they have been assigned. Initially, *E* is 0, since every individual is in a cluster of its own. At each stage the link created is the one that makes the least increase to *E*" (Upton and Cook 2008).

5 “Two types of *p*-values are available: approximately unbiased (AU) *p*-value and bootstap probability (BP) value. Multiscale bootstrap resampling is used for the calculation of AU *p*-value, which has superiority over BP value calculated by the ordinary bootstrap resampling.” (Suzuki and Shimidora 2006: 1540). AU-values are generated by bootstrap re-sampling (1,000 iterations) with replacement. An AU *p*-value >95% indicates that the respective cluster is highly supported.

6 Since the use of unclassified words and cardinal numbers are relatively uninteresting with regard to variational distinctions, the corresponding three lines of Table 3 are not subject to closer investigation in the present paper.

References

Belasubramanian, C. (2009), *Register Variation in Indian English*. Amsterdam: Benjamins.

Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: CUP.

Conrad, S. and D. Biber. (2001), *Variation in English – Multi-dimensional Studies*. Harlow: Pearson.

Garside, R. (1987), ‘The CLAWS word-tagging system’, in: R. Garside, G. Leech and G. Sampson (eds) *The Computational Analysis of English: A corpus-based approach*. London: Longman. 30-41.

Halliday, M. A. K. (1989), *Spoken and Written Language*. Oxford: OUP.

Hundt, M. and U. Gut (2012), *Mapping Unity and Diversity world-wide.* Amsterdam: Benjamins.

Leech, G. (2006), ‘New resources, or just better old ones? The holy grail of representativeness’, in: M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus Linguistics and the Web.* Amsterdam: Rodopi. 133-149.

Nelson, G., S. Wallis and B. Aarts (2002), *Exploring Natural Language. Working with the British Component of the International Corpus of English.* Amsterdam. Benjamins.

Suzuki, R. and H. Shimidora (2006), ‘Pvclust: an R package for assessing the uncertainty in hierarchical clustering’, *Bioinformatics*, 22: 1540-1542.

Upton, G and I. Cook (2008), *A Dictionary of Statistics* (2nd rev. ed.), Online version. Oxford: OUP. EISBN: 9780191726866.

Ward, J. H. (1963), ‘Hierarchical grouping to optimize an objective function’, *Journal of the American Statistical Association*, 58: 236-244.

Are word-stress variants in lexicophonetic corpora exceptional cases or regular forms?

Franck Zumstein

Université Paris Diderot

Abstract

This study of word-stress variation relies on a sub-corpus of more than 2,000 word entries in which stress variants appear. They were extracted from a computer-searchable version of John Wells' first edition of the Longman Pronunciation Dictionary (henceforth LPD1, Wells 1990). With the help of a selection of examples, I want to demonstrate that word-stress variation is the result of conflicting rules that are indicators of simplification in the phonology of British English. The theoretical framework adopted here is Lionel Guierre's study of word stress (Guierre 1979) that includes a close examination of exhaustive results automatically obtained from a computer-searchable version of Daniel Jones's twelfth edition of the English Pronouncing Dictionary (Jones 1963). More often than not, Guierre's Normal Stress Rule (hence NSR) is involved in the conflicts.[1] *A close examination of the variations shows that NSR stressing appears to be the new variant that challenges traditional stress patterns. We may call this process "NSR regularisation". Word stress variation is undoubtedly symptomatic of ongoing changes that are phonetically abrupt, as exemplified by the poll preferences in LPD, and lexically gradual since some words in identified lexical paradigms are not affected by the changes. In this study, directions of the changes are determined with the help of diachronic data extracted from pronouncing dictionaries of the 18th and 19th centuries. After such identifications of stress variations within the lexicon, further research needs to be carried out to put dictionary data to the test.*

1. Introduction

Anyone who has ever opened Wells' *Longman Pronunciation Dictionary* (hence LPD) to check the pronunciation of a two-syllable word such as *garage* in British English is confronted with no less than five variant phonemic transcriptions, including two different stress patterns: /ˈɡærɑːʒ/, /ˈɡærɑːdʒ/ or /ˈɡærɪdʒ/ with primary stress falling on the first syllable and /ɡəˈrɑːʒ/ or /ɡəˈrɑːdʒ/ with primary stress falling on the second syllable. A fairly large number of word entries are strewn with such segmental and stress free-variants in the dictionary. Preference polls appear in some entries and have been augmented with statistics that include some socio-linguistic variables in the 2008 third edition of LPD (hence LPD3). Free variations here may challenge any theory aiming at accounting for regular and predictable stress placement rules in English words according to well-identified and simple principles (Ballier 2005, Danielsson 1948, Guierre 1966a, 1966b, Fudge 1984, Podlauf 1984, Fournier 1993, 2007, Duchet 1994a, 1994b, Duchet & Zumstein 1998, Trevian 1998, 2000, Zumstein 2005). Which variants are regular? Which are exceptions? In this paper, analyses of some word-stress variations in British English such as /ɪkˈskwɪzɪt/ - /ˈekskwɪzɪt/ or /ˈæbdəmən/ - /æbˈdəʊmən/,[2] automatically extracted from a computer-searchable version of the 1990 first edition of LPD (hence LPD1), tend to support Wells' contention that:

> Some sound changes can be explained on the grounds that they lead to greater simplicity in the grammar (in the widest sense of this term, i.e. including phonology). This involved simplifying not the physical movement of the articulators but the abstract mental plan of the language which underlies our ability to speak it. There is always a pressure to remove irregularities by bringing irregular forms under the general rule. (Wells 1982)

Within Lionel Guierre's methodological framework, assignment of primary stress in polysyllabic words is the result of phonological, morpho-phonological and grapho-phonemic rules, derived from statistical evaluations of word-stress paradigms automatically detected in a computer-searchable lexico-phonetic corpus.[3] As a follow-up, this study shows that stress variants under scrutiny are products of ongoing linguistic changes because of stress-rule conflicts. My contention here is that Guierre's Normal Stress Rule is usually a regularising force when competing with minor stress rules that create the "irregular forms" which Wells alludes to.

2. Word-stress variation in this study

Well-known examples of word-stress variations include a process known as stress shift or rhythm rule (Selkirk 1984, Grabe and Warren 1995) as exemplified in the figure below.

stress clash				stress shift			
	X	**X**		**X**		X	
X	X	X		X	X	X	
X	X	X	X	X	X	X	X
tʃaɪ	ˈniːz	ˈɑː	mi	ˈtʃaɪ	niːz	ˈɑː	mi

Figure 1. Stress shift: *Chinese Army* /tʃaɪˈniːz/ + /ˈɑːmi/

The constraint, which entails a shift of the primary stress to an earlier syllable, does not lie within the word itself but beyond the word's boundaries. Other constraints, though, are word-internal constraints that account for stress displacement. In a word with an adjectival ending <-al> as in *horizontal*, for example, the presence of a pre-final consonant cluster retains the primary stress on the penultimate: /ˌhɒrɪˈzɒ**nt**əl /, but stress shifts back to the antepenultimate in words where a single consonant appears before the ending: *vertical* /ˈvɜːtɪ**k**(ə)l/. Primary stress may also shift to another syllable in a derived form through morphological derivation as in *origin* /ˈɒrɪdʒɪn/ > *original* /əˈrɪdʒən**(ə)l**/.

All examples here are constrained variations. In this study, I use the term of word-stress variation to refer to changes of primary stress allotment within a single word form, irrespective of any word-external constraint. This type of variation under scrutiny may be termed as free variation since it affects a single word form and native speakers and learners may choose one stress pattern or the other(s).

3. Theoretical framework

The theoretical framework for word-stress placement was devised by Lionel Guierre (1921-2001) at the very beginning of the 1980s. His seminal work appeared in the form of a 900-page long PhD dissertation (Guierre 1979) on English word stress. In an obituary, the authors describe Guierre's work as follows:

> Guierre's research was based essentially on the hypothesis that, contrary to the views of many phoneticians and phonologists at the time, word-stress in English could be explained by a system, relatively simple in itself, but differing from those often previously proposed. (Deschamps and O'Neal 2001)

As early as 1968, he and his team started digitizing Jones's 1963 twelfth edition of the *English Pronouncing Dictionary* (hence EPD12) in order to turn it into a computer-searchable lexico-phonetic corpus. Guierre put findings found in *Sound Pattern of English* ([hence *SPE*] Chomsky and Halle 1968) to the test of statistical results derived from automatic data retrieval. He showed that "the formal rules such as those proposed in *SPE* were not statistically validated" (Deschamps and O'Neal 2007). Guierre decided then to lay down his own word-stress principles which are best described in the following lines:

> Very soon, the formal rules which he initially formulated incorporated the essential role of spelling and morphology. In his enormous (990 page) doctoral dissertation (1979), he studied his corpus in its entirety (some 35,000 words) and demonstrated the considerable importance of isomorphism in derivation processes – neutral suffixing and neutral and non-neutral prefixing. He also evaluated the productivity of the various stress-imposing suffixes and showed that most word-stress rules reached or approached 100% regularity. In addition to this systematic analysis of English word-stress, his research demonstrated

> the reliability of the English graphic system. By applying the same methods of exhaustive analysis of his corpus he showed that a series of contextualized reading rules could explain the relations linking spelling and sounds in English. (Deschamps and O'Neal 2001)

Guierre concludes that the stress system of English is statistically driven by two major principles, which he gathers under the single name of "Normal Stress Rule" (NSR). This rule, or more precisely this statistical tendency, stipulates that most two-syllable words are stressed on the penultimate and most other polysyllabic words are stressed on the antepenultimate. He then evaluates sets of minor rules that account for violation of these two principles. Parameters, or rather combinations of parameters, that are to be taken into account in such cases are:

- syntactic category;
- stress-imposing word endings;
- isomorphism, notably via semantically transparent derivations;
- morphological structure;
- default phonological rules.

Along these lines, Deschamps (2000, 2001) sets forth his own account of word-stress variation. At the core of his demonstration are the conflicting rules, which may be "activated" when accounting for stress placements in many words.[4] On the one hand, conflicts may be resolved in favour of a single rule resulting in a single stress pattern for a single word form. On the other hand, unresolved conflicts usually produce competing stress patterns for a single lexical unit. Later analyses of a number of stress variation cases will show how antagonistic rules interact and thus produce various stress patterns that any speaker of British English, native or learner, is left to choose from.

4. The investigated lexico-phonetic corpus

One facet of Guierre's work was the compilation of lexico-phonetic searchable corpora. After having digitized EPD12 and other personal subcorpora of his own, he contacted the Longman publishing company who agreed to give him the electronic file that had been used to print LPD1. He used automatic processes to enrich the corpus with additional metadata for each word entry. The additional information includes reverse spelling, syllable count, codes for stress patterns,[5] part-of-speech information and a set of different symbols operating as syllable separators. Phonemic transcriptions are given in the same order as they appear in the dictionary, meaning that the first one is the main pronunciation,[6] defined as "the form recommended for EFL purposes" by Wells in the introduction of the dictionary.[7] Subsequent transcriptions are alternative recognized and accepted variants. American English main and variant pronunciations appear after the symbol || only if different from British English.

A rather lengthy description of the corpus content can be found in a previous paper (Zumstein 2006) and an exhaustive account is given in my study of word-stress variation (Zumstein 2007).

5. Data collection

We used this electronic corpus called 'A_Z9CD;2' to extract any word entry in which stress variants were detected with the help of a now outdated piece of software called Macintosh Programmer's Workshop (Hence MPW).[8] More than 2,000 word entries were thus extracted from this corpus and classified according to variations of stress patterns as exemplified below in the case of three-syllable words.

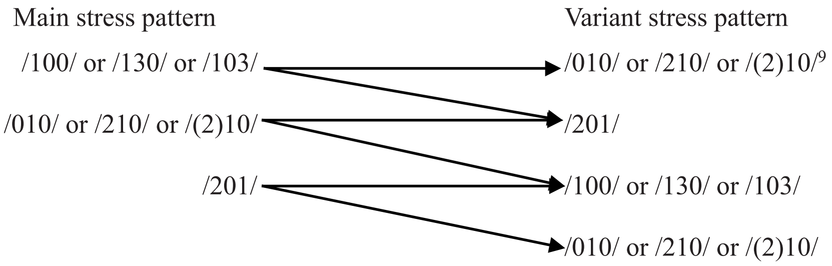


Figure 2. Three-syllable word-stress variations

Table 1. A sample of /010/ or /210/ or /(2)10/ → /100/ or /130/ or 103/ stress variations

1	2	3	4	5	6	7
preamble	N	pri ˈæm bᵊl	ˈpriː ˌ̥æm - ~s z	3	010	elbmaerp
precedent	Adj.	prɪ ˈsiːd ᵊnt	ˈpres ɪd ənt, ˈpriːs - , - əd - ~ly li	3	010	tnedecerp
prefixal	Adj.	(ˌ)priː ˈfɪks ᵊl	ˈpriː fɪks - ~ly i	3	(2)10	laxiferp
projectile	Adj. & N	prəʊ ˈdʒekt aɪᵊl	ˈprɒdʒ ekt - , - ɪkt - || prə ˈdʒekt ᵊl (*) ~s z	3	010	elitcejorp
prolative	Adj.	prəʊ ˈleɪt ɪv	ˈprəʊl ət - || proʊ ˈleɪţ ɪv	3	010	evitalorp

Table 1 above shows samples that have been extracted from one resulting table in which:

- the 1st column includes the orthographic forms of headwords;
- the 2nd column includes syntactic information;
- the 3rd column includes the main British English phonemic transcription;
- the 4th column includes variant transcriptions in British English and, after the symbol ||, the main pronunciations in American English and possible variants;
- the 5th column includes syllable counts;
- the 6th column includes stress patterns;
- the 7th and last column includes reverse spellings.

In my study of stress variation (Zumstein 2007), I give an exhaustive account of all the various stages of the data retrieval process.

6. Direction of the change

In most cases, variations of stress patterns are symptomatic of ongoing changes, so that two competing stress patterns related to the same lexical form involve an established historical stressing and a newer variant. It is then recommended to evaluate the evolution of the changes with the help of older pronunciation dictionaries. Luckily enough, 18th and 19th century orthoepists made various dictionaries in which they systematically encoded stress position in each word entry. To the best of my knowledge, the oldest of these dictionaries is Bailey's dictionary entitled *An Orthographical Dictionary Shewing both the Orthography and the Orthoepia of the English Tongue*, published in 1727. There is additional information in the long title, which indicates that the dictionary comes indeed with "Accents placed on each Word, directing to their true Pronunciation". Stress marks appear in the orthographic form of all headwords.

Of all dictionaries of the 18th century, Samuel Johnson's *Dictionary of the English Language* is probably the most famous one. Johnson relied on Bailey's work for stress patterns as he also indicated stress position in the spelling of each headword.

As for English pronunciation proper, the most successful piece of work at the time was John Walker's *Critical Pronouncing Dictionary and Expositor of the English Language*, first published in 1791, and re-published many times with changes and additions throughout the 19th century. In this dictionary, Walker used a phonetic re-spelling as a kind of pre-IPA system to account for the pronunciation of each headword, and all transcriptions of the words' pronunciations were syllabified. He actually based his work on an earlier dictionary called *A General Dictionary of the English Language*, published in 1780 by another famous lexicographer whose name is Thomas Sheridan. Appearing as a subtitle of Walker's dictionary is the statement of purpose: "One main Object of which, is, to establish a plain and permanent Standard of Pronunciation".

In a previous paper (Zumstein 2006), I showed that penultimate stress variants that can still be found for <-ate> ended verbs of three or more syllables are relics from the past and are most probably disappearing variants. Analyses of compiled data indeed indicate that in 18th century English, this type of verbs had two stressings: antepenultimate stress if the ending <-ate> is preceded by a single consonant (hence <-Cate>), and penultimate stress if the ending is preceded by a consonant cluster (at least two consecutive consonants, [hence <-C_2ate>]). Both stress patterns were in complementary distribution. Very few other words such as *administrate* /ədˈmɪnɪ**str**eɪt/ and *scintillate* /ˈsɪntɪleɪt/, stressed on the antepenultimate, stood as exceptions at the time. Yet they were "avant-gardistes" exceptions because most 3^+-syllable <-ate> ended verbs are now stressed on the antepenultimate. The pre-final consonant cluster is not functional anymore as a stress imposing device.

Apparently, a verb like *commentate* was in use at the end of the 18th century according to dated quotations found in the online version of the second edition of the *Oxford English Dictionary* (hence OED2), but it is marked as "rare" when used as a synonym of the much more frequent verb *comment*. Yet, the verb *commentate* ("commentate, v.". *OED Online*. Oxford University Press. http://www.oed.com.rproxy.sc.univ-paris-diderot.fr/view/Entry/37062) acquired a new meaning in the second half of the century with the advent of new broadcasting technologies and new media as shown in sense entry 3 below.

(1) OED entry of *commentate*
commentate, v. - Pronunciation: /ˈkɒmənteɪt/ - Etymology: A modern formation, apparently < commentator n. **1.** trans. = comment v. 2. rare. 1794 T. J. Mathias Pursuits of Lit. i. 222, Shakespeare..Almost eat up by commentating zeal. 1818 H. J. Todd Johnson's Dict. Eng. Lang., Commentate, to annotate, to write notes upon [citing Mathias]. 1864 Spectator 31 Dec. 1500, Refined prelates of the Medicean type-the men who commentated not Fathers, but only poets. 1883 Athenæum 9 June 725/1, Men who..cannot speak a word of the languages they criticize and commentate. **2.** intr. = comment v. 3 - 5. rare. 1828 Scott Jrnl. 3 Feb. (1941) 183, I corrected proofs and commentated. 1859 Sat. Rev. 8 98/1, The Commentator..had been taken in by one as competent..to commentate as himself. 1861 G. H. Kingsley in F. Galton Vacation Tourists & Trav. 1860 123, The deer, indeed, rather like the sheep..and a flock scampering about three or four miles off is instantly seen and commentated on by them. **3. To deliver an oral commentary, esp. upon politics or sport; to act as a commentator** (see commentator n. 2b, 2c). Freq. const. on. 1951 H. Nicolson Diary 26 Oct. (1968) 211, [I have] given three commentaries... William Clark and McKenzie also commentate. Labour leads during the night. 1977 H. Douglas-Home Birdman (1978) iii. 44, I lifted her gently to show the eggs to the children, commentating all the time. 1979 Washington Post 27 May d4, A former college gymnastics coach who now commentates on the sport for ABC-TV. 1984 Times 23

> July 8/2, James Burke. commentated on the original moon landing. (Simpson, J.& Weiner, E. (eds) 1989)

It is then not surprising that the verb does not appear in the successive editions of Walker's and Johnson's dictionaries at the end of the 18th century and throughout the 19th century. In LPD1, this verb has no penultimate stressing since it is a very recent coinage.

Finally, the penultimate stressing of a verb like *concentrate* does not appear as a variant in LPD1, though it was recorded in 18th century pronunciation dictionaries.

Today, all <-ate>-ended verbs are thus classified in a uniform /-100/ stress paradigm,[10] based on the sole identification of the graphemic ending <-ate>. A purely synchronic posture would thus lead us to consider the penultimate stressing either as an irregular variant, or an emergent new stressing. However, a diachronic stance allows us to resolve this question since the /-010/ stress pattern is just a leftover of a past regularity.

7. NSR regularisation

7.1 Latin Stress rule versus NSR

In the verbs described in Section 6, shifting the stress back from the penultimate to the antepenultimate also means that the stress pattern of the verbs complies to Guierre's NSR. Being the major rule of English stress, it does not come as a surprise that many patterns are regularized on a final /-100/ stress pattern in polysyllabic words of more than 2 syllables. Other words exhibit such regularization as in the following examples of Latinate learned words.

Table 2. Latinate learned words with penultimate stressed tense vowels. Variation /010/ → /100/

Words	LPD1		OED	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
chimera	kaɪˈmɪərə	ˈkɪm*ə*rə	/kɪˈmɪərə/	*
congener	kənˈdʒiːnə	ˈkɒn*d*ʒɪnə	/ˈkɒndʒɪnə(r)/	*
cerebrum	səˈriːbrəm	ˈserəbrəm	/ˈsɛrɪbrəm/	*
gravamen	grəˈveɪmen	ˈgrævəmən	/grəˈveɪmɛn/	*
subsidence	səbˈsaɪdᵊnᵗs	ˈsʌbsɪd*ə*nᵗs	/səbˈsʌɪdns/	/ˈsʌbsɨdns/

Table 3. Latinate learned words with penultimate stressed tense vowels. Variation /100/ → /010/

Words	LPD1		OED	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
ibidem	ˈɪbɪdem	ɪˈbaɪdem	/ɪˈbaɪdɛm/	/ˈɪbɪdɛm/
patina	ˈpætɪnə	pəˈtiːnə	/ˈpatᵻnə/	/pəˈtinə/ (American English stressing)
acumen	ˈækjʊmən	əˈkjuːmen	/ˈakjʉmən/	/əˈkjumən/ (American English stressing)
abdomen	ˈæbdəmən	æbˈdəʊmen	/ˈabdəmən/	/abˈdəʊmən/
resida	ˈresɪdə	rɪˈsiːdə	/ˈrɛsᵻdə/	/rᵻˈsiːdə/

Words in Table 2 and Table 3 are relevant examples showing that English stress placement may correlate to vowel quantity in the penultimate syllable, meaning that tense vowels tend to retain stress on the syllable in which they are pronounced, a tendency usually considered to be inherited from Latin. This Latin Stress Rule, much commented in the literature of generative phonology, is summed up in (2) below.

(2) Phonological rule: weight-to-stress principle
Heavy syllables (= /-(C_n)V_{TENSE}-/ or /-VC_2-/) are stressed.[11]

Yet, variants with antepenultimate stress indicate that many of those Latinate learned words are being anglicised and conform to the NSR. Variation in Table 2 shows that historical stress patterns are still in use as recommended pronunciations in LPD1, but new NSR variants have made their way into the language. In the case of the word *subsidence* ("subsidence, n.". *OED Online*. OEP. http://www.oed.com.rproxy.sc.univ-paris-diderot.fr/view/Entry/193001), the authors of the OED2 note that

> subsidence, n. **Pronunciation:** Brit. /səbˈsʌɪdns/ , /ˈsʌbsᵻdns/ , U.S. /səbˈsaɪdns/ , /ˈsəbsədns/ (…) N.E.D. (1914) gives both pronunciations. The traditional pronunciation, which is given first, places the stress on the second syllable, but the form with the stress on the first syllable, under the influence of residence n.1 and subsidy n., is recorded from the early 20th cent. and has gained in currency (Simpson, J.& Weiner, E. (eds) 1989).

For this word, Wells conducted a poll preference in Britain in 1988-1989 from which he concludes that 47% of the respondents prefer the /010/ pattern and a majority, 53%, prefers the /100/ stressing. Yet, Wells still records the former as the main pronunciation. The contradiction here between the results of the preference poll and the author's recommended pronunciation means that the scope of this phonological rule is more and more reduced in English phonology. The study of <-ate>-ended words in Section 6 and other Latinate adjectives ending in <-al> below also tend to support this assertion.[12]

A fair amount of Latinate adjectives ending in <-al> have seen their primary stress being retracted to an earlier antepenultimate syllable in variant pronunciations.

Table 4. Latinate adjectives ending in <-al>. Variations /-010/ → /-100/ & /-100/ → /-010/

Words	**LPD1**		**OED**	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
doctrinal	dɒkˈtraɪnᵊl	ˈdɒktrɪnᵊl	ˈdɒktrɪnəl	dɒkˈtraɪnəl
intestinal	ɪnˈtestɪnᵊl	ˌɪnteˈstaɪnᵊl	ɪnˈtɛstɪnəl	*
coronal	ˈkɒrənᵊl	kəˈrəʊnᵊl	kɒˈrəʊnəl	ˈkɒrənəl
communal	ˈkɒmjʊnᵊl	kəˈmjuːnᵊl	ˈkɒmjʉnl	kəˈmjuːnl
gingival	dʒɪnˈdʒaɪvᵊl	ˈdʒɪn*d*ʒɪvᵊl	dʒɪnˈdʒaɪvəl	*

In the case of *vicinal*, LPD1 only retains the NSR antepenultimate stress pattern, but OED2 still records the historical /010/ pattern as a variant. The adjective *marital* is stressed on the first syllable in both dictionaries but Wells mentions that it used to be pronounced /məˈraɪtᵊl/. Surprisingly, Johnson, Sheridan and Walker have only a /100/ stress pattern for this word in their respective 18th century dictionaries. In the mid-19th century, Thomas Wright (1852) also retains this pronunciation of the adjective in his pronouncing dictionary. Wells most certainly makes reference to former editions of Jones's EPD, in which the author introduced this /010/ variant pronunciation. Jones and the successive editors of the dictionary kept record of this stress pattern to abandon it from the 1997 15th edition on. In the same vein, it seems that Johnson introduced the /010/ stressing of the adjective *doctrinal* ("doctrinal, adj. and n.". OED Online. March 2014. OUP. http://www.oed.com.rproxy.sc.univ-paris-diderot.fr/view/Entry/56317) in his dictionary as recounted by the authors of OED2.

> doctrinal, adj. and n. **Pronunciation:** /ˈdɒktrɪnəl/ /dɒkˈtraɪnəl/. The historical pronunciation, < Latin *doctrīˈnālis*, French and Middle English *doctriˈnal*, is *ˈdoctrinal* (so Bailey, Todd); *docˈtrīnal* (Johnson) passes over the actual Latin, French and Middle English

> words, to reach the ulterior *doctrīna* (…). (Simpson, J. & Weiner, E. (eds) 1989)

Today's NSR variant of *doctrinal* is actually the true etymological pronunciation that originated from an iambic reversal applied on the Latin base *doctrīˈnālis* (> *doctriˈnal* > *ˈdoctrinal*). Johnson linked the adjective to a different Latin etymon, *doctrīna*, in some sort of diachronic phonological hypercorrection. Jones may have taken the same path when introducing the penultimate stressing of *marital*.

In Guierre's theoretical framework, the adjectival termination<-al> is considered as a stress-imposing ending along the following lines.[13]

(3) Stress-imposing adjectival <-al>
a) NSR if <-al> is preceded by a single consonant: *ˈfederal, muˈnicipal, ˈnominal, reˈciprocal, suˈbliminal, oˈriginal (< ˈorigin), heˈretical (< ˈheretic), maˈniacal (< ˈmaniac), ˈhexagon (< heˈxagonal)*, etc.
b) Penultimate stress if <-al> is preceded by a functional consonant cluster: *paˈrental* (<*ˈparent*), *ˌhoriˈzontal* (< *hoˈrizon*), *maˈgistral*, *caˈdastral*, *ˌconsoˈnantal* (< *ˈconsonant*), etc.

The effect of NSR regularisation in contemporary English is that the Latinate adjectives in Table 4 have a penultimate stress, which is considered as exceptional, or more precisely pertaining to a different stress rule.[14] Their recommended antepenultimate variant has made them join the larger paradigm of adjectives whose stress pattern follow rule (3a). Nevertheless, numerous examples of Latinate <-al>-ended adjectives, many of which end in <-idal> and <-ival>, do not comply to such NSR regularisation via variant stress patterns, as exemplified in Table 5 below.

Table 5. Latinate <-al>-ended adjective with a single /010/ stress pattern

Words	LPD1		OED	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
genitival	ˌdʒenəˈtaɪvᵊl	*	dʒɛnɪˈtaɪvəl	*
relatival	ˌreləˈtaɪvᵊl	*	ˌrɛləˈtʌɪvl	*
fungicidal	ˌfʌŋgɪˈsaɪdəl	*	fʌndʒɪˈsaɪdəl	*
suicidal	ˌsuː‿ɪˈsaɪdᵊl	*	s(j)uːɪˈsaɪdəl	*
decretal	diˈkriːtᵊl	*	dɪˈkriːtəl	*
sinusoidal	ˌsaɪnəˈsɔɪdᵊl	*	saɪnəˈsɔɪdəl	*

Resistance to change shows that the process of NSR regularisation is lexically gradual.

7.2 French Stress Rule versus NSR

There are many other instances of NSR regularisation, especially among words with final syllables bearing primary stress. Many words borrowed from French belong to this category as shown in Table 6 below.

Table 6. French words ending with <-ine>. Variation /201/ → /100/ & /100/ → /201/

Words	LPD1		OED	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
guillotine	ˈɡɪləti:n	ˌɡɪləˈti:n	ˈɡɪləti:n	ɡɪləˈti:n
magarine	ˌmɑ:dʒəˈri:n	ˈmɑ:dʒəri:n	ˌmɑ:dʒəˈri:n	ˌmɑ:ɡəˈri:n
magazine	ˌmæɡəˈzi:n	ˈmæɡəzi:n	ˌmaɡəˈzi:n	ˈmaɡəzi:n
quarantine	ˈkwɒrənti:n	*	ˈkwɒrənti:n	*
gelatine	ˈdʒeləti:n	ˌdʒeləˈti:n	ˈdʒɛlətɪn	-(ˈ)i:n

Massive borrowings from French occurred at different stages in the history of the English language and allowed an "embedded foreign stress system" to settle within a Germanic structure (Fournier 1991, Zumstein 2007).

Further new borrowings have been posited in such an embedded structure with the particularity of being stressed on the last syllable, that is "à la française". While some of the borrowings are irremediably stuck in the embedded structure,[15] many have "escaped" from it since their primary stress is now retracted to earlier syllables according to the NSR (*ˈbutton*, *ˈbenefit*, *aˈcademy*, *caˈlamity*, etc.). Others are moving out of the embedded structure and, before settling in the NSR-ruled macro-structure, are going through connecting zones with stress variations as results of conflicting rules. The following figure represents the migration within the English system.

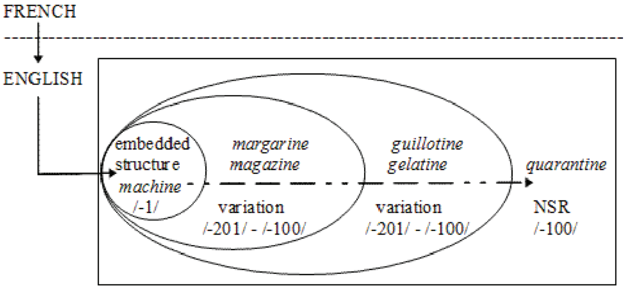


Figure 3. From final stress to antepenultimate stress

NSR stress assimilation shows that:

> Uncertainty prevails until the foreignness of the adopted word is adjusted in order to fit into the phonetic arrangements and accentual system of English. Some words remain partially or permanently in a zone of incomplete adaptation. (Fowler 1996)

As mentioned in the last sentence of the above citation, this process of regularisation is not lexically abrupt since other word forms do not have NSR stress variants.

7.3 Morpho-phonological rules versus NSR

Within our theoretical framework, prefixes are stress imposing in monocategorial verb forms as exemplified below:

(4) Stress rule for monocategorial prefixed verb: stress pattern /01-/ or /201-/.[16]
deˈtermine, conˈsider, ˌrepreˈsent, acˈcept, beˈcome, etc.

Along the same lines, the verbs *contribute*, *attribute* and *distribute* do not have primary stress on the syllable of their prefixes, nor do the adjectives *exquisite* and *recondite*. NSR treatment, which produces antepenultimate variant stressings for these words is thus blind to their morphological structure since primary stress falls on the prefixes in such cases.

Poll panel preferences were conducted by Wells for the words *contribute*, *distribute* and *exquisite* and the results in LPD1 show that in each case, just over one fourth of the respondents are in favour of NSR stressing. In the third edition, LPD3, the results for *contribute* are updated and NSR receives the favour of more than one-third of the respondents.[17] Both adjectives and the three verbs are not regarded as prefixed words anymore. Common graphemic endings unify those verbs on one side, and the adjectives on the other. Consequently, the verbs *attribute*, *contribute* and *distribute* may join the paradigm of NSR-stressed <-ute>-ended 3-syllable words such as *ˈsubstitute, ˈconstitute, ˈprosecute*, etc. As for the adjectives, they may thus integrate the paradigm of 3-syllable words ending in <-ite> and stressed on the antepenultimate such as *opposite, ˈfavourite, ˈrequisite*, etc.). In both cases, speakers use a reading rule, that is to say the identification of an NSR stress-imposing graphemic ending.[18]

Table 7. Three syllable <-ute>-ended verbs and <-ite>-ended adjectives. Variation /010/ → /100/ & /100/ → /010/

Words	LPD1		OED	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
contibute	kənˈtrɪbjuːt	ˈkɒntrɪbjuːt	/kənˈtrɪbjuːt/	*
attribute	əˈtrɪbjuːt	ˈætrɪbjuːt	/əˈtrɪbjuːt/	*
distribute	dɪsˈtrɪbjuːt	ˈdɪstrɪbjuːt	/dɪˈstrɪbjuːt/	*
exquisite	ekˈskwɪzɪt	ˈekskwɪzɪt	/ˈɛkskwɪzɪt/	/ɪkˈskwɪzɪt/
recondite	ˈrekəndaɪt	rɪˈkɒndaɪt	/ˈrɛk(ə)ndʌɪt/	/rɨˈkɒndʌɪt/

7.4 NSR and other conflicting rules

Another major rule that Guierre puts forward in his framework is isomorphism via stress-neutral suffixation in semantically transparent derivation processes as exemplified in (5).

(5) Neutral suffix <-able>. Stress Preservation Rule.
acˈcountable (< *acˈcount*), *beˈlievable* (< *beˈlieve*), *diˈstinguishable* (<*diˈstinguish*), etc.

This neutral suffix may be contrasted with the stress-imposing suffix <-ic>:

(6) <-ic>-ended words: stress pattern /-01/
ˌacroˈbatic (< *ˈacrobat*), *geˈnetic* (< *ˈgene*), *sylˈlabic* (<*ˈsyllable*), etc.

Yet, in LPD1, exceptional stress patterns are recorded as variants of otherwise regular penultimate <-ic> stress patterns of certain words, which are listed in Table 8.

In the <-ic>-ended words, the exceptional /-100/ stress pattern resembles NSR stress placement, but this is most certainly coincidental.[19] In fact, the antepenultimate stressing of those derived forms is to be related to the penultimate stressing of the deriving forms: *stomach* /ˈstʌmək/, *choler* /ˈkɒlər/ and *psoriasis* /sɔːˈraɪəsɪs/. The derivational process does not take into account the stress effect of the termination <-ic>, so that stress isomorphism marks the semantic transparency of the suffixation in such cases. Such an exceptional process for <-ic>-ended words may certainly be confined to usage in the medical sphere where these words are most frequently pronounced. Conversely, the pre-antepenultimate stress variant of *despicable* is not in line with the stress pattern of the verb *despise*, /dɪˈspaɪz/. Duchet and Fournier (1988) comment on this particular example and conclude that NSR stress retraction has operated on its stressable portion (i.e. <despic->) because the modality in the underlying predicative relation has changed from “may” to “must”.[20]

Table 8. Stress variants of words ending in <-ic> (variation /-10/ → /-100/ & /-100/ → /-10/) and in <-able>

Words	**LPD1**		**OED**	
	Main stress pattern	**Variant stress pattern**	**Main stress pattern**	**Variant stress pattern**
stomachic	stəˈmækɪk	ˈstʌməkɪk	/stəʊˈmækɪk/	*
choleric	ˈkɒlərɪk	kɒˈlerɪk	/ˈkɒlərɪk/	*
psoriatic	ˌsɔːriˈætɪk	sɔːˈraɪətɪk	/ˌsɒrɪˈatɪk/	*
despicable	dɪˈspɪkəbəl	ˈdespɪkəbəl	/ˈdɛspɪkəb(ə)l/	*

8. Conclusion

Word-stress variations as analysed within Guierre’s theoretical framework in this paper are traces of ongoing suprasegmental changes. It seems that the “grammar” of oral English tends to be simplified by way of regularising variants whose stress patterns conform to the requirement of a major rule of the stress system of English: the Normal Stress Rule.
The changes here are:

> - phonetically abrupt as they are noticeable since Wells’ poll preferences, which punctuate some word entries in his dictionary, show that native speakers are aware of the variations;
> - lexically gradual since conflicting rules producing variations within a class of words do not concern all the eligible words of the class.

In this respect, word-stress variation as described here fits in a theory of lexical diffusion (Wang 1977). Father work needs to be done within this framework, taking into account Bybee’s recommendations with regard to word frequency and context of use.

> My goal is to demonstrate that word frequency and context frequency are factors that can affect variation and should be taken into account in future studies of phonological variation and change. (Bybee 2002)

Diachronic data is also necessary in such a project so as to indicate the direction of the changes. I am thus engaged in a lengthy task of digitizing 18th and 19th century pronouncing dictionaries so as to turn them into searchable corpora and eventually web-searchable databases. As a final step, I most certainly need to go a step further in the study of stress variation, that is to say, beyond the dictionary, and collect sufficient empirical oral data with identified suprasegmental variations so as to put the type of ongoing changes studied here to the test.

Notes

[1] The Normal Stress Rule imposes primary stress on the antepenultimate of words of three or more syllables.

[2] *Exquisite* and *abdomen* respectively.

[3] Or combinations of those rules.

[4] Deschamps prefers here to use the French term “logiques accentuelles” to echo the title of his paper and show that variation forces one to explain the reasons that lead to such conflicts.

[5] /1/ for syllable bearing primary stress, /2/ for syllable bearing secondary stress, /3/ for syllables bearing and /0/ for unstressed syllable.

[6] It is possible to search for segments according to syllable ranking.

[7] See LPD3, p.xvii.

[8] The MPW tool zone website is no longer maintained since MPW has been replaced by XCODE (https://developer.apple.com/xcode/), but an MPW fan has kept an info webpage where it is still possible to download MPW: http://www.geek-central.gen.nz/MPW/intro.html.

[9] Read (2) as an optional secondary stress.

[10] Read /-100/ as the final stress pattern from the stressed syllable on, as in *concentrate* /**100**/, *communicate* /0**100**/ and *disambiguate* /20**100**/. Primary stress is on the antepenultimate syllable of the words.

[11] Syllables in which the vowel is tensed or followed by a consonant cluster are called “heavy syllables”, hence the weight-to-stress principle.

[12] Interestingly enough, Walker (1797) wrote the following lines in the long introduction of his dictionary: “The first general rule that may be laid down, is, that when words come to us whole from the Greek or Latin, the same accent ought to be preserved as in the original. Thus *horizon*, *sonorous*, *decorum*, *dictator*, *gladiator*, *mediator*, *delator*, *spectator*, *adulator*, &c. preserve the penultimate accent of the original ; and yet the antepenultimate tendency of our language has placed the accent on the first syllable of *orator*, *senator*, *auditor*, *cicatrix*, *plethora*, &c. in opposition to the Latin pronunciation of these words and would have

infallibly done the same by *abdomen*, *bitumen*, and *acumen*, if the learned had not stepped in to rescue these classical words from the invasion of the Gothic accent, and to preserve the stress inviolably on the second syllable." The last three examples are now stressed on the antepenultimate, a stress pattern recorded as the main pronunciation in LPD1. The word *bitumen* has no penultimate variant.

13 See also Trevian (2003).

14 The Latin Stress Rule.

15 Examples here are words like *caˈnal*, *maˈchine*, *poˈlice*, *faˈtigue*, etc.

16 Read /01-/ and /201-/ as the initial stress pattern of a word from the beginning of the word until the syllable bearing primary stress. Primary stress does not fall on the syllable, or any of the syllables, of the prefix.

17 These are the results of an online preference poll conducted in April-June 2007.

18 It may be argued that <-ute>, <-ate> and <-ite> are suffixes rather than mere graphemic endings. I think that, when adopting a diachronic point of view, these forms may be considered as suffixes but, synchronically, speakers have easier access to the words' graphemic forms than their morphological structure which is, for most words, now opaque: *agreg + ate? *elabor + ate? *recond + ite? *persec + ute? Only a few of them have transparent morphological structures with existing deriving forms: oppos(e) + ite, oxygen + ate, orchestr(a) + ate, etc.

19 When two stress rules may be applicable to the same word form and produce the same stress pattern, we may term this process as "rule conspiration".

20 When decribing a person, the adjective despicable means that the person is to be despised, not that (s)he may be despised.

References

Ballier, N. (2005), 'De l'exception à la régulation des exceptions pour la phonologie de l'anglais: métarègles ?', in: I. Vilkou-Poustovaïa (ed) *Faits de Langues. L'Exception, entre les Théories Linguistiques et l'Expérience*, 25. Paris:Ophrys. 245-253.

Bybee, J. (2002), 'Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change', *Language Variation and Change*, 14: 261-290. Cambridge: CUP.

Chomsky, N. and M. Halle (1968), *The Sound Pattern of English*. New York: Harper & Row.

Danielsson, B. (1948), *Studies on the Accentuation of Polysyllabic Latin, Greek,* *and Romance Loan-Words in English, with special Reference to Those Ending in -able, -ate, -ator, -ible, -ic, -ical and -ize.* Stockholm: Almqvist and Wiskell.

Deschamps, A. (1994), *De l'Écrit à l'Oral et de l'Oral à l'Écrit: Phonétique et Orthographe de l'Anglais.* Paris: Ophrys.

Deschamps, A. (2000), 'La Logique des variantes accentuelles de l'anglais', in: *Point d'Interrogation. Phonétique et Phonologie de l'Anglais.* Université de Pau et des Pays de l'Adour. Publications de l'Université de Pau. 93-107.

Deschamps, A. (2001), 'Stress patterns, rules and variants: Can stress variation be accounted for?', in: A.-M. Harmat, L. Hewson, J. Durand and D. Philips (eds) *Anglophonia, A French Journal of English Studies*, 9, Toulouse: Caliban-Presses universitaires du Mirail. 41-57.

Deschamps, A. and M. O'Neal (2007), 'Obituary, Lionel Guierre', *Language Sciences*, 29: 492-495

Duchet, J.-L. (1994a), *Code de l'Anglais Oral*, 2nd edition. Paris. Ophrys.

Duchet, J.-L. (1994b), 'Éléments pour une histoire de l'accentuation lexicale en anglais', *Études Anglaises*, 47: 161-170.

Duchet, J.-L. and J.-M. Fournier (1988), 'Isomorphisme et productivité dans l'accentuation et la prononciation des mots dérivés anglais', in: P. Larreya and J. Humbley (eds) *Actes du 4e Colloques d'Avril sur l'Anglais Oral*. Université de Paris-Nord: CELDA, diffusion APLV.

Duchet, J.-L. and F. Zumstein (1998), 'Étude de la variation accentuelle des mots terminés en <-etive>, <-itive>, <-otive>, <-utive> et <-ative> fondée sur un corpus lexico-phonetique informatisé', in: N. Ballier, A. Deschamps and J.-L. Duchet (eds) *Actes du 9e Colloques d'Avril sur l'Anglais Oral*. Université de Paris 13: APLV. 33-48.

Fournier, J.-M. (1991), 'Voix ethnique – Voix des ethnies: le cas de l'anglais', *Voix Ethniques - Ethnic Voices*. Tours. Publications du GRAAT.

Fournier J.-M. (1993), 'Motivation savante et prononciation des adjectifs en -ic en anglais contemporain', *Faits de langues. Motivation et iconicité*, 1. Paris: Presses Universitaires de France.

Fournier, J.-M. (2007), 'From a Latin syllable-driven stress system to a Romance vs Germanic morphology-driven dynamics', *English Phonology, Language Sciences*, 29: 218-236.

Fowler, H. W. (1996), *The New Fowler's Modern English Usage*, 3rd edition. Oxford: OUP.

Fudge, E. (1984), *English Word Stress*. London: Allen and Unwin.

Grabe, E. and P. Warren (1995), 'Stress shift: do speakers do it or do listeners hear it', in: B. Connell and A. Arvanitti (eds) *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, 4. Cambridge: CUP. 95-110.

Guierre, L. (1966a), 'Éléments pour une étude linguistique de l'accentuation en anglais', *Les Langues Modernes*, 1: 161-170.

Guierre, L. (1966b), 'Traitement automatique des langues: un codage des mots anglais', *Études de Linguistique Appliquée*, 4: 48-63.

Guierre, L. (1979), *L'Accentuation en Anglais Contemporain, Éléments pour une Synthèse*, Paris: Département de Recherches Linguistiques, Institut d'anglais Charles 5. Université de Paris 7 – Denis Diderot.

Jones, D. (1963), *English Pronouncing Dictionary*, 12th edition. London: Dent-Dutton.

Podlauf, I. (1984), *English Word-Stress: A Theory of Word-Stress Patterns in Englis*h. Oxford and New York: Pergamon Press.

Selkirk, E. (1984), *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press.

Simpson, J. and E. Weiner (1989), *The Oxford English Dictionary*, 2nd edition. OED Online. OUP. Available online at www.dictionary.oed.com

Trevian, I. (1998), 'Règles et variations accentuelles: vers une troisième voie entre isomorphisme et rétraction germanique ? Le cas de la terminaison -atory', in: N. Ballier, A. Deschamps and J.-L. Duchet (eds) *Actes du 9e Colloques d'Avril sur l'Anglais Oral*. Université de Paris 13: APLV. 49-65.

Trevian, I. (2000), 'Variants and phonetic changes in lexemes with irregular realisations: Is the English language overcoming its phonological conflicts?', in: P. Busuttil (ed) *Points d'Interrogation, Phonétique et Phonologie de l'Anglais*. Pau: Université de Pau. 73-90.

Trevian, I. (2003), *Morphoaccentologie et Processus d'Affixation de l'Anglais.* Bern: Peter Lang.

Walker, J. (1797), *A Critical Pronouncing Dictionary and Expositor of the English Language*, 2nd edition. London. G.G. and J. Robinson, Paternoster-Row; and T. Cadell, junior, and W. Davies in the Strand.

Wang, W. S.-Y. and C.-C. Cheng (1977), 'Implementation of phonological change: the Shuang-Feng Chinese case', in: W. S-Y. Wang (ed) *The Lexicon in Phonological Change*. The Hague: Mouton. 141-158.

Wells, J. (1982), *Accents of English 1, an Introduction*. Cambridge: CUP.

Wells, J. (1990), *Longman Pronunciation Dictionary*, 1st edition. Harlow: Longman.

Wells, J. (2008), *Longman Pronunciation Dictionary*, 3rd edition. Harlow: Longman.

Wright, T. (1852-1856), *The Universal Pronouncing Dictionary, and general Expositor of the English language*. 5 volumes. London: The London and New York Publishing Company, Limited.

Zumstein, F. (2005), 'De la règle à l'exception et de l'exception à la règle: le cas des variantes accentuelles et segmentales du lexique de l'anglais contemporain', in: I. Vilkou-Poustovaïa (ed) *Faits de Langues. L'Exception, entre les Théories Linguistiques et l'Expérience*, 25. Paris: Ophrys. 255-258.

Zumstein, F. (2006), 'The contribution of computer-searchable diachronic corpora to the study of word-stress variation', in: R. Facchinetti and M. Rissanen (eds) *Linguistic Insights 31: Corpus-based Studies of Diachronic*

English, Bern: Peter Lang. 171-196.

Zumstein, F. (2007), *Variation Accentuelle, Variation Phonétique: Étude Systématique Fondée sur des Corpus Lexico-phonétiques Informatisés Anglais*, PhD dissertation, unpublished. Université de Poitiers.

Part 2. Specialist corpora

Relative clauses in Philippine English: a diachronic perspective

Peter Collins, Xinyue Yao* and Ariane Borlongan***

*University of New South Wales, Sydney
**De La Salle University, Manila

Abstract

This paper explores aspects of diachronic change in a non-native variety of English, Philippine English. It uses the Philippine section of the International Corpus of English (sampling period early 1990s) and a new corpus 'Phil-Brown', parallel in its design and sampling date (early 1960s) to the LOB and Brown corpora. Comparison is made between PhilE and the two super-varieties, British and American English, drawing from the pioneering work by Leech et al. (2009) on grammatical change in contemporary written English. The study focuses on relative clauses, more particularly that*-relatives and* wh-*relatives. It was found that Philippine English has followed the two super-varieties in experiencing a decline in* wh-*relatives and an increase in* that*-relatives, but differs from them in the rapidity with which the changes have occurred, reflecting an attempt to approximate the patterns of its 'colonial parent', American English. When we compare the rates of change for the frequencies of* that*-relatives and* wh-*relatives across the genres we find more indications of Philippine English progressively aligning itself with American English, and in turn further evidence that the linguistic orientation of Philippine English remains predominantly exonormative.*[1]

1. Introduction

This paper examines recent changes involving relative clauses in Philippine English ('PhilE'), with comparisons made with the findings of studies of relative clauses in American English ('AmE') and British English ('BrE'). The study upon which the paper is based represents a new development in corpus linguistics, a marriage of short-term corpus-based diachronic study as epitomised in the work of Leech et al. (2009), and corpus-based grammatical research on World Englishes (much of it published in *World Englishes* and *English World-Wide*). The diachronic study of PhilE has recently become possible with the near-completion of a new corpus ('Phil-Brown') in a project directed by Ariane Borlongan at De La Salle University. The period of time defined by the sampling period for Phil-Brown (the late 1950s – early 1960s) and that for a second corpus that we shall also be using in this study, ICE-Phil, the Philippines component of the *International Corpus of English* (comprising texts sampled in the early 1990s), covers most of the period of time over which there has been general recognition of the existence of PhilE as a World English. Important historical onset dates were the end of World War II and the Declaration of Independence in 1946. Since then the exonormative orientation of English in the Philippines towards its parent variety, AmE, has begun to give way to a new acceptance of that variety simply referred to as PhilE, whose status as an official language is today shared with Filipino.

How far has PhilE developed beyond the external norm-dependence that characterized English in the Philippines in the decades following 1898, when the United States was granted authority over the Philippines and AmE began to be disseminated in the public school system? The second half of the twentieth century, following the end of World War II and the Philippines' attainment of Independence in 1946, saw a gradual lessening in this exonormativity. However, the extent to which PhilE today has undergone endonormative stabilization is a matter of some debate. According to Schneider (2007), the endonormativity of PhilE is merely incipient. Schneider points to the perseverance of complaints about 'errors' and 'incorrect' uses of English. More recently, however, Borlongan (2011) has argued that PhilE is in fact well established, and that political events that suggest an increasing sense of Philippine independence from the United States have had a contributory role to play in this process (events such as the rejection of the Military bases Agreement in 1991, and the recall by the Philippines of its humanitarian contingent from Iraq in 2007). Borlongan also cites empirical evidence of the development of increasingly PhilE-positive attitudes in the Philippines, as in his own study of younger Filipinos (Borlongan 2009) which suggests that their customary embarrassment over the use of PhilE has largely given way to acceptance of it as a legitimate variety. In this paper we shall consider whether there is evidence of exonormativity or endonormativity in the relativization strategies of PhilE, and furthermore whether these are undergoing change.

The regional and stylistic dimensions of relativization have been quite extensively studied. Most studies of the interrelationship between relativization and regional varieties have focused either on BrE (see Quirk 1957, Ball 1996, Hoffman 2005), or AmE (see Kikai et al. 1987, Guy and Bailey 1995). Useful sources of corpus-based frequency information on both varieties are Leech et al. (2009) and Biber et al. (1999). Relativization has also been investigated in New Zealand English by Sigley (1997), in PhilE by Coronel (2010), and in a range of New Englishes by Gut and Coronel (2012). Studies that have considered the interrelationship between relativization and genre include Guy and Bailey (1995), Ball (1996), Tottie (1997), Sigley (1997), Biber et al. (1999), and Hoffmann (2005). The present study both complements earlier studies, in its concern with both regional and stylistic dimensions of relativization, and extends them, with its pioneering diachronic analysis of relativization in a New English.

This study focuses on the main overt relativizers in standard varieties of English (*that*, *which*, *who*, *whom* and *whose*). Relative adverbs (*where*, *when*, *why*, etc.) were excluded from the study on the grounds of their very low frequencies, while zero relatives were excluded because of the difficulty of automatically extracting tokens.[2] The choice of relativizer is subject to such factors as the questions of whether the relative clause is integrated or supplementary, and whether the antecedent is human or non-human: see Section 3 below.

The frequencies of the relativizers are known to be sensitive to differences of region and genre, and Leech et al. (2009) have shown that they have been undergoing significant diachronic change in recent decades. Leech et al.'s research concentrates on the three decades between the early 1960s and the early 1990s, and the four corpora on which the bulk of their findings are based are those that are commonly referred to as the 'Brown family', whose nuclear members are as follows: Brown (representing written AmE of the early 1960s), LOB (1960s BrE), Frown (1990s AmE) and FLOB (1990s BrE). The strength of these corpora lies in their parallel design, with virtually the same size and selection of texts and genres, and paired sampling dates. By comparing frequencies in LOB and FLOB with those in Brown and Frown respectively, Leech et al. are able to present the percentage rises and falls for the relativizers in BrE and AmE, and in various genres (by calculating the difference in the frequencies between the 1960s and 1990s corpora as a proportion of those for the 1960s corpora). In conjunction with insights gained from a consideration of some of the semantic and grammatical factors that constrain relativizer choice, such findings enable Leech et al. to identify various broad patterns of grammatical change across the two supervarieties and to shed light on the linguistic and socio-historical factors that have led to these changes. Accordingly, we shall make systematic comparisons between our findings for PhilE and those of Leech et al. for the supervarieties.

In Section 2 we present the composition of the two PhilE corpora used in this study, Phil-Brown and ICE-Phil. In Section 3 we describe the grammatical properties of relative clauses. Section 4 presents findings for the diachronic variation of the relativizers in PhilE, drawing comparisons with Leech et al.'s (2009) findings for BrE and AmE. Section 5 examines the frequencies of the relativizers in three macro-genres: press, learned and fictional writing. Finally, Section 6 is devoted to concluding remarks.

2. Phil-Brown and ICE-Phil

As noted above, the data-sources for the present study were the two PhilE corpora, Phil-Brown and ICE-Phil. Phil-Brown, which is designed to be a Philippine counterpart of Brown and LOB, is now approximately two-thirds complete (with some 674,000 words collected so far). Table 1 presents the composition of Brown, the founding member of the Brown family of corpora (under the 'target' column), and that of Phil-Brown as at the time of writing this paper (under the 'current' column).[3] The fifteen text types of Brown are subdivided into four 'macro-genres', three of which (press, learned writing and fiction) were selected for this study on the grounds that they span much of the range of situational and linguistic variation in written English. Learned writing is produced by and for specialists, and thus contrasts with press and fiction, which have in common that they tend to be popular rather than specialized, but differ in that press is factually-oriented and fiction is used to depict imaginary situations.

The category of general prose was excluded for two reasons: it is stylistically heterogeneous, comprising a broad array of specialist and non-specialist non-fiction; and no Philippine texts have yet been collected for the categories of 'skills, trades, and hobbies' or popular lore, which together account for 41% of the general prose category in Brown.[4]

Table 1. Design of Phil-Brown

Genre	**Category**	**Content**	**No. of words (approx)**	
			Target	**Current**
Press	A	Press reportage	88,000	28,000
	B	Press editorials	54,000	30,000
	C	Press reviews	34,000	59,000
General prose	D	Religion	34,000	130,000
	E	Skills, trades, hobbies	72,000	0
	F	Popular lore	96,000	0
	G	Belles lettres, biographies	150,000	60,000
	H	Miscellaneous	60,000	118,000
Learned	J	Learned	160,000	83,000
Fiction	K	General fiction	58,000	91,000
	L	Mystery/detective fiction	48,000	0
	M	Science fiction	12,000	0
	N	Adventure and western	58,000	61,000
	P	Romance and love story	58,000	14,000
	R	Humour	18,000	0

The text categories of Phil-Brown overlap considerably with the written categories of ICE-Phil. Table 2 represents the set of text categories from the two corpora that were selected and matched for analysing stylistic variation in the use of relative clauses. This selection of text categories is employed as the basis of the empirical analysis in Sections 4 and 5.

Table 2. Selected text categories from Phil-Brown and ICE-Phil

	Phil-Brown		**ICE-Phil**	
	Category	**No. of words (approx)**	**Category**	**No. of words (approx)**
Press	A+B+C	118,000	W2C+W2E	68,000
Learned	J	83,000	W2A	88,000
Fiction	K+N+P	167,000	W2F	48,000

3. Grammatical properties of relative clauses

In this section we explain and illustrate the grammatical distinctions that form the basis of our analysis of relative clauses in Section 4 and Section 5. Relative clauses are so called because they are related to an antecedent, which determines the interpretation of the anaphoric element that they contain within their structure. In the case of '*wh*-relatives' (whose initial element is constituted by, or – in the case of complex relativizers – includes, one of the relativizers *who*, *whom*, *whose*, *which*, etc.) the anaphoric element is overt, as in (1) where the pronoun *which* is anaphoric to *things*.

By contrast, there is no overt anaphoric element in the relative clauses exemplified in (2-3). This is clearly so in the case of the 'zero-relative' in (3) in which there is a gap representing the understood object of *had*, but also in the case of (2) on the analysis – which we accept – of *that* as a subordinator rather than relative pronoun (in accordance with which analysis the relative clause contains a gap representing the understood object of *despised*).

(1) The only things **which existed only in myth, fable and fairy tale a thousand years ago,** now are as common to us in the 20th century as camels were in Biblical times. (Phil-Brown Press)
(2) It meant doing all the things **that Emma despised**. (Phil-Brown Fiction)
(3) They exchanged stories and looked for things **they had in common**. (ICE-Phil Fiction)

In the following sections we discuss and exemplify three sets of grammatical properties that were subjected to quantitative analysis in the study: the integrated versus supplementary distinction, the distinction between simple versus complex relativizers, and the types of antecedents that relativizers may have.

3.1 Integrated versus supplementary relatives

The distinction between 'integrated' and 'supplementary' relatives overlaps with, but is not completely coextensive with, the more familiar distinction between 'restrictive' and 'non-restrictive' relatives.[5] By contrast with integrated relatives, which present their content as integral to the meaning of the clause, supplementary relatives are non-integrated, presenting their content as a separate, parenthetical piece of information that could be omitted without affecting the sentence's meaning. The relatives in (1-3) above are all of the integrated type and serve, as integrated relatives typically do, as modifier of the antecedent (in these examples, *things*). The vast majority of integrated relative clauses serve semantically to restrict the denotation of the antecedent expression: in (3), for example, the relative clause restricts the denotation of things to the subset of 'things they had in common'. There are, however, exceptions, non-prototypical cases such as (4) below, where an integrated relative does not serve this semantically restrictive function. In (4), the *that* relative does not serve to restrict the denotation of *table* by distinguishing the table in question from other tables, but nevertheless presents information about the table that is integral to the sentence.

(4) a small pig was grunting and pilling at its rope tied to one leg of the table **that shook against the low concrete wall partition** (Phil-Brown Fiction)

Consider (5), which contains a supplementary relative clause.

(5) The Internet, **which is turning the world into a true global village**, is the nightmare of authoritarian regimes. (ICE-Phil Press)

In (5) the antecedent for *which is turning the world into a true global village* is not the noun *Internet* but rather the NP *the Internet*, which refers to a unique institution: the relative clause does not serve to identify the referent of the antecedent, merely supplying extra information about the referent.

While an integrated relative usually functions as an NP modifier, accordingly forming a syntactic constituent with the antecedent NP head, supplementary relatives have a less determinate syntactic structure as a result of their loose incorporation within the sentence. In the clearest cases in written English commas are used, or occasionally dashes or brackets, to mark off supplementary relatives. Unfortunately, however, punctuation is not always a reliable clue, given its unsystematic use in written language. Consider example (6), where the commas that one might have expected are presented in square brackets. Given that the antecedent is a proper noun and thus fully defined, an integrated interpretation is not possible.

(6) The fertility of the alluvial soil in the valley replenished by yearly inundation of the Cagayan river [,] **which deposits its rich deposits over the vast Cagayan region** [,] gives the area a tremendous productivity potential (Phil-Brown Press)

In view of such inconsistencies, punctuation could not be relied upon in the task of (manually) distinguishing between integrated and supplementary relatives in our analysis. Useful clues to distinguish the two categories were the strong tendency for *that*-relatives to be integrated and *which*-relatives to be supplementary, and for antecedents in the form of clauses or proper NPs to be associated with supplementary relatives. Despite the availability of detailed accounts of the different meaning and uses of integrated and supplementary relatives found in the descriptive grammars of English, once we embark on a systematic corpus-based investigation we find tokens which are ambivalent between the two types (and certainly not helped by the punctuational inconsistencies described above). Such cases, which account for less than 5% of all tokens, are excluded from the percentages presented in the discussion of integrated and supplementary relatives in Section 4.4 below.

3.2 Simple versus complex relativizers

Relativizers may be ‘simple’ (e.g. *which* in (1) above) or ‘complex’ (i.e. relative phrases consisting of a relative word in combination with one or more other elements). The major types of complex relativizer are as presented below.

(i) PP

This type, formed with *which* or *whom* as head of the relative phrase, is exemplified in (7). The derivation of the relative clause from its canonical counterpart *the Thomasites worked*, involving the relativisation and fronting of the PP *under the circumstances*, is on some accounts described metaphorically as involving ‘pied-piping’ of the preposition along with its NP complement. An alternative, less formal, version would have just *whom* as relative phrase and the preposition left ‘stranded’: *which the Thomasites worked under.* A corpus example featuring preposition stranding is (8), which would sound inappropriately formal with pied-piping (*out of which they ate*):

(7) The circumstances **under which the Thomasites worked** were indeed very difficult (ICE-Phil Learned)

(8) They, folks and friends, had just started to put away the lunch boxes **which they ate out of** while they waited in the crematory site. (ICE-Phil Fiction)

(ii) NP with embedded PP

In this type a relativized PP is embedded in a larger NP, as in (9):

(9) A recent report on spices contained in FAO Bulletin No. 8/1960, **parts of which are quoted below** will help to awaken us to the possibilities (Phil-Brown Learned)

(iii) NP with *whose* as determiner

In this case *whose* functions as determiner in a relative NP, as in (10):

(10) I looked at Pedro, **whose eyes seemed to be popping out**, so white they were (Phil-Brown Fiction)

(iv) PP or NP with *which* as determiner

This type, found only in supplementary relatives, involves an NP containing determinative *which*, which serves as complement within a relative PP as in (11), or as an element in clause structure as in (12).

(11) The table can also burn, **in which case, a drastic change takes place**. (ICE-Phil Learned)

(12) But herein lies a maddening corollary to their findings, **one which the old villagers understood equally well**: (ICE-Phil Fiction)

3.3 Relativizers and their antecedents

A noun serving as antecedent for a relative clause may be human or non-human. As observed by Quirk et al. (1985: 1246-1250), Sigley (1997: 217-220) and Huddleston and Pullum (2002: 1048-1050), *who* and *whom* are used exclusively with human antecedents, as in (13) and (14), while *whose* is used mostly with human antecedents as in (15), and more rarely with non-human antecedents as in (16).

(13) We do not look for a messiah **who will vindicate a metaphysical tale from which we can then draw every truth**. (ICE-Phil Learned)

(14) Nong Tomas, **for whom a sled had been obtained**, was very sick. (Phil-Brown Fiction)

(15) The identity of the man **whose guilt or innocence shall be the business of this tribunal to establish** makes our task a grave one. (Phil-Brown Fiction)

(16) The foreshore area is public domain, **whose disposal in this instance has been determined by bayside municipalities and a contracting firm**. (Phil-Brown Press)

By contrast, *which* and *that* are used mostly with non-human antecedents, as in (1-2) above. An example of relative *that* with a human antecedent – observed by Sigley (1997) to be found most commonly in informal spoken language – is (17):

(17) I remembered the women in Rome, the only woman in that foreign city **that I had known** with just a bit beyond casual intimacy, while I was now sketching the hips of the matron before me, … (Phil-Brown Fiction)

4. Frequency findings

In this section we present findings for the diachronic variation of the relativizers in PhilE, drawing comparisons with Leech et al.'s (2009) findings for BrE and AmE. The retrieval method used in the study was to search for all tokens of *that*, *which*, *who*, *whom* and *whose*, and then manually weed out all irrelevant (that is, non-relative) tokens.

4.1 *That* versus *wh*

Let us begin by comparing the frequencies for *that*, and the four *wh*-relatives as a set, in BrE, AmE and PhilE, in the 1960s corpora and the 1990s corpora, presented as percentages in Figure 1. Due to the lack of a general prose section in the PhilE corpora, the frequencies in this figure were derived by calculating an average figure for their per million word (pmw) frequencies in the press, learned and fiction genres. Note that the comparisons presented should not be taken to suggest that our attention was restricted – since it certainly was not – purely to interchangeable relativizers (that is, cases where relativizer *that* can be substituted for a *wh*-relativizer, and vice versa). As Figure 1 indicates, the frequency of the *wh*-relatives outstrips that of the *that* relatives, in all three varieties during both time periods. Nevertheless, *that* is relatively more strongly represented, and *wh* more weakly. In Table 3 percentages are used – in this case quite differently from those in Figure 1 – to quantify the diachronic changes undergone by *that* and the *wh*-relatives between the 1960s and 1990s (the difference between the 1960s and 1990s frequencies calculated as a percentage of the former).

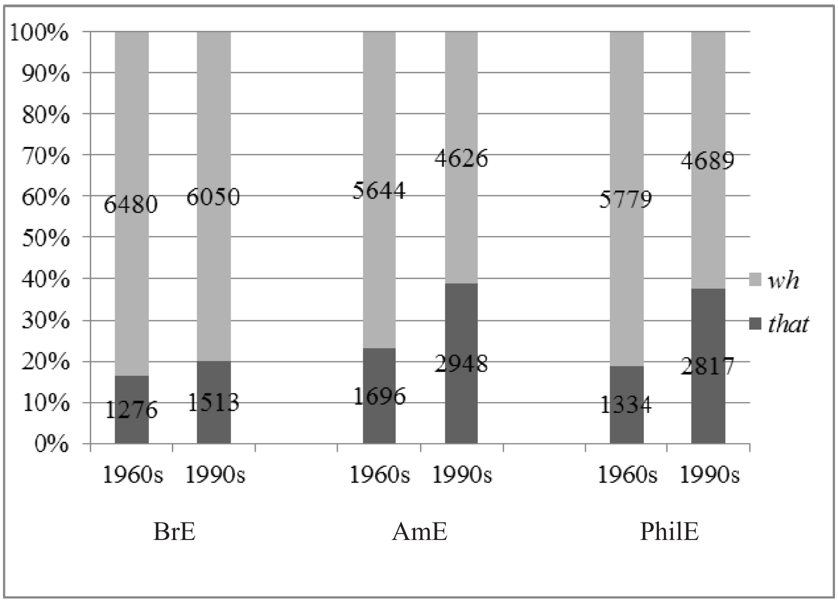


Figure 1. Changes in the relative frequencies of *that* vs *wh*-relativizers in PhilE, AmE and BrE (based on averages of their pmw frequencies in press, learned and fiction). AmE and BrE frequencies are from Leech et al. (2009: 308-310).

Table 3. Frequency changes for *that* and *wh*-relatives in the 1960s and 1990s corpora

	BrE	**AmE**	**PhilE**
that	+18.6%	+73.8%	+111.2%
wh	-6.6%	-18.0%	-18.9%

Some interesting patterns emerge if we consider the frequencies presented in Figure 1 in conjunction with the percentage differences in Table 3. The frequencies of *that* in the 1960s are similar in the case of PhilE (1334) and BrE (1276), while AmE has a considerably higher frequency (1696). AmE retains its ascendancy in the 1990s, with a 73.8% increase to 2948, and BrE displays its traditional conservatism with a modest increase of only 18.6%, to 1513. PhilE makes up considerable ground on its colonial parent with a massive increase of 111.2% to 2817.

By contrast with the consistent story of rising frequencies across the three varieties with *that*, *wh*-relatives are in decline, both collectively as a set and, as we shall see below, individually. Once again it would appear that PhilE is patterning more with AmE than BrE: AmE and PhilE have very similar frequencies in the 1960s and 1990s, and therefore a very similar rate of decline (-18.0% for AmE, and -18.9% for PhilE). The more conservative of the two supervarieties, BrE, begins and ends the three decade period with a higher frequency – unsurprisingly given the formality associations – of the *wh*-relatives and suffers only a comparatively mild decline (-6.6%).

The divergence between BrE and AmE and convergence between AmE and PhilE is also apparent if we compare the ratios of *wh* to *that* relatives in the 1960s and 1990s corpora. We find a quite similar dominance of *wh*-relatives over *that* relatives in the 1960s in AmE and BrE (3.3:1 for Brown; 5.1:1 for LOB). However, when we compare this situation with that in the 1990s we find that the gap between BrE and AmE has widened considerably: in BrE *that* relatives have increased slightly and *wh*-relatives have decreased slightly, reducing the ratio to 4.0:1, but in AmE the increase and decrease have been sharper, reducing the ratio dramatically to 1.6:1. As for PhilE, we find that it starts from a position in the 1960s (with a *that* vs *wh* ratio of 4.3:1) that is more like that in the more conservative supervariety, BrE (5.1:1), than it is to AmE (3.3:1). However, the rate of change evidenced in PhilE has been rapid as in AmE rather than modest as in BrE, resulting in the ratio reducing to 1.7:1 in the 1990s, which is far closer to the 1990s AmE ratio of 1.6:1 than it is to the BrE ratio of 4.0:1. As noted above, given the historical fact that AmE is the colonial ‘parent’ of PhilE, it seems plausible to suggest that PhilE has been playing catch up to AmE.

A significant factor in the complementary diachronic trends of *that*-expansion and *wh*-attrition that we have identified in this section is undoubtedly ‘colloquialization’, the process responsible for the widespread informalization of hitherto more formal genres in contemporary English. Relative *that* is an overtly colloquial grammatical feature that has been found to be more common in spoken than written English. For instance Biber et al. (1999: 609-612) found relative *that* to be more frequent in conversation than zero or the *wh*-relativizers, while Coronel (2010) found it to be the most popular of the relative markers in the spoken section of ICE-Phil. By contrast *wh*-relatives are a non-colloquial feature (found more commonly in formal written style and subject to prescriptive rules), so it is plausible to suggest they have been negatively impacted by the forces of colloquialization.

It may be noted that the diachronic fortunes of relative *that* bear similarities to those of another overtly colloquial feature, the quasi-modals (as investigated in a recent study by Collins (ms) of diachronic change in modality in PhilE, AmE and BrE). Like relative *that*, the quasi-modals are undergoing a rapid rise, one in which AmE is leading the way in sheer frequency terms. As with relative *that*, so with the quasi-modals: PhilE appears to be responding to external American norms, with a rate of increase which is even higher than that in AmE and which suggests that it is attempting to bridge the frequency gap with its parent variety.

A further factor in the rise of relative *that* is prescriptivism. Since this involves the alternation between *that* and the individual *wh*-relativizer *which*, discussion is reserved for the following section.

4.2 The individual relativizers

Let us now focus on the individual relativizers. Figure 2 displays the 1960s and 1990s frequencies for *that*, *which*, *who*, *whom*, and *whose* in the two Phil-Brown and ICE-Phil subcorpora. Unfortunately no comparable frequencies are available for the individual *wh*-relativizers in AmE and BrE.[6]

Arguably the most interesting comparison here is between *which* and *that*, which are typically felicitously interchangeable in integrated relatives. However, there is a continuing tradition – particularly in the USA – of proscribing *which* and prescribing *that* in usage guides, pedagogical grammars, and grammar checking software. The rationale behind the tradition, which was first recorded as a rule in Fowler (1926) and subsequently popularized in Strunk and White's (1959) enormously influential usage guide, appears to be that if *that* is impossible in non-restrictive relatives, in order to maintain symmetry *which* should be impossible in restrictives. The increasing favoring of *that* is undoubtedly due, at least in part, to this tradition. In a recent paper, Szmrecsanyi (2012) uses a set of prescriptivism-related predictors to account for the striking decline of restrictive *which* in late 20th century AmE, concluding that the *which*/*that* rule represents one of the rare cases in which prescriptivists have managed to initiate a change in language usage.

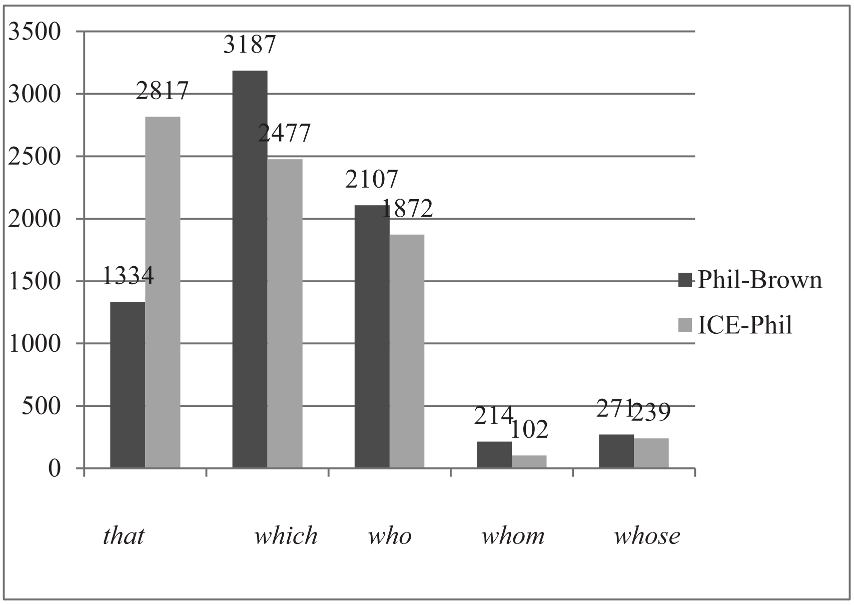


Figure 2. Frequencies for relativizers in Phil-Brown and ICE-Phil (based on averages of their pmw frequencies in press, learned and fiction)

The ICE-Phil frequencies for *that* and *which* in Figure 2 show that in the 1990s PhilE *that* (2817) has overtaken *which* (2477). Biber et al.'s (1999) bar graph frequencies for the supervarieties combined suggest that *that* is yet to overtake *which*, but one suspects that this results from a combination of American advancement (attested for *that*: see Section 4.1 above) and British conservatism (attested for the *wh* items as a set: see Section 4.1). It is likely, then, that the relationship between relative *that* and *which* in contemporary PhilE has more in common with this relationship in contemporary AmE than in contemporary BrE.

As Figure 2 shows, while *which* is the most frequently occurring relativizer in our Phil-Brown subcorpus (where the order is *which* > *who* > *that* > *whose* > *whom*), over the three decades that separate Phil-Brown and ICE-Phil, *that* has leapfrogged both *who* and *which* into first position, the ordering in the ICE-Phil subcorpus being *that* > *which* > *who* > *whose* > *whom*. The latter ordering is confirmed by Coronel's (2010: 35) findings for ICE-Phil.[7] The percentage changes for the individual relativizers presented in Table 4 show that the reordering results from a combination of diachronic shifts: a massive increase in the frequency of *that* (+111.2%) together with a mild fall in the frequency of the highest frequency *wh*-relativizer, *which* (-22.3%). The extent of this decline – stronger than the -9.4% decline of relative *which* noted by Leech et al. (2009: 309) in BrE, but milder than that of -34.4% in AmE – suggests that PhilE writers may be subject to similar prescriptive pressures to their American counterparts (see also Section 4.1 above).

Is it possible that, in addition to colloquialization, one factor in the rise of *that* might be an increasing acceptance in its use with *human* antecedents, enabling it to take over some of the territory previously occupied by the *wh*-relativizers? The answer would seem to be 'no'. The numbers for this use of *that* are very small (33 in Phil-Brown and 39 in ICE-Phil), leading us to conclude that, while acceptance of human *that* may be increasing, it represents only a very minor contributing factor in the rising popularity of *that.*

Table 4. Percentage changes for the individual relativizers in PhilE (derived from Figure 2)

that	*which*	*who*	*whom*	*whose*
+111.2%	-22.3%	-11.2%	-52.3%	-11.8%

The percentage change figures in Table 4 confirm that of the four *wh*-relativizers it is *whom*, the lowest frequency item, that has suffered the sharpest fall (-52.3%).[8] The fortunes of *who* and *whom* should probably be considered together in view of their similarly strong association with human antecedents (in Coronel's 2010 study, 98% of *who* tokens in writing, and 97.3% in speech, took a human antecedent, while for *whom* the corresponding percentages were 89.0% and 97.6%). Why is *whom* declining at a far stronger rate (-52.3%) than *who* (-11.2%)? As Figure 2 indicates, there is a considerable frequency difference between them, with *who* outstripping *whom* by a ratio of 2107:214, or 9.8:1, in the 1960s, and by 1872:102, or 18.4:1, in the 1990s. Could it be that another factor in the relative mildness of the decline of *who* is its capacity to be used in object function? Closer investigation of this hypothesis suggested that the answer to this question must be negative: objective relative *who* is in fact extremely rare in the PhilE corpora, with only two tokens in Phil-Brown, and one in ICE-Phil (see below):

(18) First there is the Beaux-Arts quartet itself, a first-rate group of individual artists upon **who** critics and the press on four continents have lavished praise (Phil-Brown Press)

(19) and he would raise his hands and meet the other kid **who**, it seemed to him he could not really beat (Phil-Brown Fiction)

(20) He was too drunk even to think, much more speculate on the appearance of the two fighters **who**, from the fringe of the lawn, he could still guess was Ben, the tallest in their gang, though both young men were tall. (ICE-Phil Fiction)

Finally, *whose*, despite being a low frequency item like *whom*, has suffered only a mild decline (-11.8%). Perhaps the greater resilience that it displays by comparison with *whom* can be attributed to its capacity to take both human and non-human antecedents (see Section 3 above) and to the absence of the type of formality associations that we find with *whom*.

4.3 Genre variation

Table 5 presents changes in the frequencies of the relativizers in the three genres discussed in Section 3 above. In this table, as in subsequent tables, asterisks placed next to a numerical quantity indicate degrees of statistical significance as determined by the log-likelihood test (absence of an asterisk = 'non-significant'; * = 'significant at the level $p < 0.05$ (log-likelihood > 3.84)'; ** = 'significant at the level $p < 0.01$ (log-likelihood > 6.63)'; *** = 'significant at the level $p < 0.001$ (log-likelihood > 10.83)').

Table 5. Change in frequencies of the relativizers across genres in BrE, AmE and PhilE. AmE and BrE frequencies are from Leech et al. (2009: 308-310).

			Press	Learned	Fiction
that	BrE	1960s/1990s	1123/1259	968/1566	1738/1714
		% diff	+12.1%	+61.8% **	-1.4%
	AmE	1960s/1990s	2017/3131	1382/3442	1689/2271
		% diff	+55.3% ***	+149.0% ***	+34.4% ***
	PhilE	1960s/1990s	1036/2261	2008/3176	959/3013
		% diff	+118.2%***	+58.3%***	+214.2%***
wh	BrE	1960s/1990s	7758/6959	7548/7612	4135/3578
		% diff	-10.3% **	+0.9%	-13.5% **
	AmE	1960s/1990s	6574/5850	6658/4895	3701/3132
		% diff	-11.0% **	-26.5% ***	-15.4% ***
	PhilE	1960s/1990s	7429/4833	5547/4963	4362/4272
		% diff	-34.9%***	-10.5%	-2.1%

In BrE and AmE of the 1960s the genre in which relative *that* is most poorly represented is learned writing (with a pmw frequency of 968 in British learned writing and 1382 in American). Learned writing is, as Leech et al. (2009: 228) observe, the most 'formally informative' variety, and thus one in which a colloquial feature such as relative *that* might be considered out of place by some writers of this period. By the 1990s, however, the situation has changed: the colloquialization of the learned genre, along with continuing prescriptive censure of the use of *which* in integrated relatives (or "*which*-hunting"), has seen a substantial increase in the frequency of relative *that* (+61.8% in British learned writing, and +149.0% in American), one considerably stronger than that encountered in press and fiction. Fiction, the most speech-like of the genres, is at the other extreme, with an already robust frequency for relative *that* in the 1960s and undergoing milder diachronic change (+34.4% in American fiction, and a non-significant -1.4% in British fiction) than in the other two genres. The British and American results for press are in-between. The results for relative *that* in PhilE appear perplexing at first blush. The learned writing and fiction genres are again at the extremes, but their positions are reversed, with learned writing evidencing a large pmw frequency in the 1960s (2008) accompanied by the mildest increase (+58.3%), and fiction a very low 1960s frequency (959) accompanied by a very strong rise (+214.2%). A possible explanation for the findings is the allegedly stylistically neutral nature of PhilE: according to Gonzales (1997, 2004), writers continue to lack sensitivity to the conventions of writing in standard native varieties, writing in much the same way that they speak.

Interestingly, however, the changes that have happened in PhilE writing in the years since the 1960s have brought it more into line with British and American writing, and not more so with the latter. The 1990s frequency for relative *that* in PhilE learned writing (3176) is similar to that for 1990s AmE learned writing (3442), while that for 1990s PhilE fiction (3013) makes it more similar to AmE fiction than it was in the 1960s.

As for the PhilE press genre, why does its openness to relative *that* stop short of allowing it to acquire a frequency commensurate with that of AmE press (2261 in PhilE press, versus 3131 in AmE)? An explanation similar to that advanced by Collins et al. (ms) for some of their results may be relevant. Observing a stronger-than-expected endorsement of modals in PhilE press, Collins et al. interpret this as reflecting a conservatism driven by press writers' predominant concern with intra-national issues.

Turning to the *wh*-relativizers, we find that the percentage changes presented in Table 6 are generally less spectacular than those for *that*. In the case of AmE there is a degree of complementarity between the results for *that* and those for *wh*: learned writing has both the strongest percentage increase for *that* (149.0%) across the three genres and the strongest percentage decrease for *wh* (-26.5%), while for press and fiction the comparatively milder increase for *that* (+55.3% and +34.4% respectively) is complemented by a milder decrease for *wh* (-11.0% and -15.4% respectively). The diachronic variations in BrE are generally less significant than those in AmE. The BrE genres are not only consistently less accommodating to the advance of colloquial *that* than those in AmE, but at the same time consistently more conservatively resistant than AmE to the decline of the *wh*-relativizers in all three genres. In PhilE the decline of *wh* across the genres is mild and fails to complement the large-scale increase in the frequency of *that*. In the case of press and learned writing, the result of the diachronic changes is that the 1990s PhilE frequencies (4833 in press, and 4963 in learned) are aligned more closely with those for AmE (5850; 4895) than with those for BrE (6959; 7612).

Closer insights into the relationship between PhilE and AmE emerge from an inspection of the *wh* versus *that* ratios in Table 6.

Table 6. *Wh* vs *that* ratios across genres in AmE, BrE and PhilE. Bold figures represent statistically significant differences over time ($p < 0.05$ on the log-likelihood test).

		Press	**Learned**	**Fiction**
AmE	1960s	**3.3:1**	**4.8:1**	**2.2:1**
	1990s	**1.9:1**	**1.4:1**	**1.4:1**
BrE	1960s	6.9:1	7.8:1	2.4:1
	1990s	5.5:1	4.9:1	2.1:1
PhilE	1960s	**7.2:1**	**2.8:1**	**4.6:1**
	1990s	**2.1:1**	**1.6:1**	**1.4:1**

What is striking in this table is the closeness of the parallels between PhilE and AmE in the 1990s subcorpora, despite considerable differences between the varieties in the 1960s. In press the 1960s difference between AmE (3.3:1) and PhilE (7.2:1) is levelled to 1.9:1 and 2.1:1 respectively; in learned writing the difference between 4.8:1 and 2.8:1 is reduced to 1.4:1 and 1.6:1; and in fiction the difference between 2.2:1 and 4.6:1 is reduced to 1.4:1 in both cases. The ratios in Table 6 also serve to clearly differentiate AmE and PhilE on the one hand from BrE on the other, the latter's non-significant ratios consistently evidencing considerably stronger conservative support for *wh*-relativizers over *that* than AmE and PhilE in all three genres. A further finding to emerge from Table 6 is that consistently in the 1990s subcorpora, across all three varieties, the *wh* vs *that* ratios for press outstrip those for learned writing, and those for learned writing outstrip those for fiction. Further research into the frequency differences between colloquial and non-colloquial grammatical alternants is required to ascertain whether the genre-related differences identified here for relative clauses are part of a more general picture.

Table 7 compares the frequency changes of individual *wh*-relativizers across the three genres in PhilE. Unfortunately comparable information for BrE and AmE is not available.

Table 7. Change in frequencies of the *wh*-relativizers (pmw) in PhilE

		Press	**Learned**	**Fiction**
which	1960s/1990s	3430/2202	4160/3723	1971/1507
	% diff	-35.8%***	-10.5%	-23.5%*
who	1960s/1990s	3328/2335	1028/990	1965/2291
	% diff	-29.8%***	-3.7%	+16.6%
whom	1960s/1990s	289/59	120/80	234/166
	% diff	-79.6%***	-33.3%	-29.1%
whose	1960s/1990s	382/236	239/171	192/310
	% diff	-38.2%	-28.5%	+61.5%

Several notable findings emerge. *Whom*, whose vulnerability to decline derives from its low frequency (see Section 4.1 above) and the formality associated with its common use in complex relatives (see Section 4.5 below), displays the strongest decline in all genres. Whereas all four *wh*-items are in decline in press and learned writing, there is heterogeneity in fiction (*which* and *whom* in decline, and *who* and *whose* on the rise, the latter perhaps related to increasing personalization). This finding might reflect the stylistically heterogeneous nature of fiction, with the increasing informalization of the genre due in large part to the increasing popularity of dialogue, but variability in the degrees of formality in non-dialogic fiction. The most significant declines are consistently recorded in press (which as we have noted above also differentiates itself elsewhere from the other genres in PhilE). This finding, taken in conjunction with the fact that, as we have seen, *that* has risen substantially in PhilE press, suggests that this genre in PhilE has undergone a substantial degree of informalization. Finally, for reasons that are unclear to us, *who* and *whose* record a rise in fiction, bucking the declining trend that is found elsewhere for the *wh*-relativizers.

4.4 Integrated versus supplementary relatives

As Table 8 indicates, integrated relatives enjoy an average percentage increase in frequency of +8.9% between the 1960s and the 1990s, while at the same time supplementary relatives have declined at a rate of -4.1%.

Table 8. Percentage changes from the 1960s to the 1990s for integrated and supplementary relatives in PhilE

	Press	**Learned**	**Fiction**	**Average**
Integrated	-12.2%	+7.8%	+31.0%***	+8.9%
Supplementary	-20.1%*	-6.8%	+14.5%	-4.1%

One factor in these developments may be colloquialization. Several studies have shown that integrated relatives are comparatively more popular in speech than writing. In Sigley's (1997) study of New Zealand English, his written data contained 79.1% restrictive relatives, his spoken data 88.5%. Coronel's (2010) findings for PhilE were 84.4% for written PhilE and 90.1% for spoken. The effects of this situation are reflected in the findings for the three written macro-genres in PhilE. In fiction, the most 'speech-friendly' of the genres, integrated relatives have enjoyed a highly significant rise (+31.0%), compared to the milder rise for supplementary relatives (+14.5%). In learned writing, integrated relatives have risen mildly (+7.8%) but at the same time supplementaries are in mild decline (-6.8%). In press, the genre which, as we have noted above, is most affected by the decline of *wh*-relatives, the decline of integrated relatives (-12.2%) is milder than that for supplementaries (-20.1%).

Consider finally the percentage changes for relativizers in integrated and supplementary relatives, presented in Table 9. Integrated *that* enjoys a massive increase (124.9%), but this is merely a by-product of the fact that relative *that* is generally rising sharply in frequency. Integrated *which* has suffered a decline, undoubtedly due in part to prescriptive censure of it (see Section 4.2 above), while supplementary *which* has barely changed. With *who* there has been a decline for both the integrated and supplementary types, slightly milder for integrated. In many ways, *whom* and *whose* are more extreme cases than *which* and *who* respectively.

Table 9. Percentage changes for relativizers in integrated and supplementary relative clauses in PhilE (based on averages of pmw frequencies for press, learned and fiction)

	that	*which*	*who*	*whom*	*whose*
Integrated	+124.9%	-36.3%	-9.7%	-68.7%	-2.3%
Supplementary	-13.9%	-2.5%	-13.4%	-25.7%	-28.7%

4.5 Complex relativizers

Table 10 presents the percentages of simple versus complex *which*-relatives and *whom*-relatives in Phil-Brown and ICE-Phil. *Whose* is not included because there is no simple counterpart to complex relativizers with *whose*. Against the general backdrop of a rapid decline for the somewhat old-fashioned and formal relativizer *whom*, complex relativizers with *whom* are furthermore faring better (-18.9%) than their simple counterparts (-71.4%), probably for the same reason.

Table 10. Percentage changes for simple versus complex relatives in PhilE

		Press	**Learned**	**Fiction**	**Average**
which	Simple	-24.7%**	-5.7%	-14.0%	-14.8%
	Complex	-77.8%***	-25.1%	-51.0%*	-51.3%
whom	Simple	-65.9%	-81.7%	-66.7%	-71.4%
	Complex	-85.7%	+13.3%	+15.7%	-18.9%

As Table 10 also indicates, complex relativizers with *which* are, predictably, declining less rapidly in the traditionally formal and conservative genre of learned writing (-25.1)% than they are in press (-77.8%) and fiction (-51.0%). That the decline should be more significant in press than fiction may seem surprising. However we have already seen (in Section 4.2 above) that the *wh*-relativizers generally are declining markedly in PhilE press (and at the same time relative *that* increasing strongly). In the case of complex relativizers with *whom*, fiction and learned writing have actually enjoyed a mild frequency increase (+15.7% and +13.3% respectively). Again, however, it is the genre of press that behaves differently, recording a large decline of -85.7%.

Table 11 presents a comparison of the diachronic changes in complex PP relatives involving 'pied-piping' in PhilE and BrE.[9] The greater decline evidenced in PhilE than in BrE is not surprising given the tendency for BrE to exhibit conservatism in grammatical change and, conversely, for PhilE to align itself more closely with the less-conservative AmE. The association of relativizer complexity with formality is also reflected in the results for learned writing, whose rate of decline in both PhilE and BrE is smaller than for the other two genres. The most significant difference between PhilE and BrE is to be found in the press results: in the case of BrE the result for press (-32.3%) is almost identical to that for fiction (-32.1%), but for PhilE – as in the case of complex relativisation with *which* and *whom* generally, as discussed above – pied-piped relatives have undergone a particularly sharp decline (-79.5%).

Table 11. Complex (pied-piped) relatives (pmw) in PhilE and BrE

	Press	**Learned**	**Fiction**	**Average**
PhilE	-79.5%***	-23.4%	-38.4%	-57.8%
BrE	-32.3% **	-1.3%	-32.1% **	-21.9%

5. Conclusion

In this paper we have availed ourselves of the unique opportunity to pursue a corpus-based examination of recent diachronic variation in a New English, PhilE, that is afforded by the (near-)completion of the 'Brown family' corpus Phil-Brown. Comparison of the 1960s written data from this corpus with 1990s written data from parallel categories of the Philippines component of the *International Corpus of English* collection has enabled us to measure recent changes that have occurred over a three-decade period. In turn this has enabled us to explore a question that has occupied the attention of scholars of PhilE, that of the continuing norm-dependence of PhilE on AmE.

Previous studies have shown that relative clauses, the grammatical category we investigate here, are subject to extensive regional and stylistic variation. More recently Leech et al. (2009) have documented – inter alia – the rise of *that* relatives at the expense of *wh*-relatives in contemporary BrE and AmE, suggesting that socio-historical factors such as colloquialization and prescriptivism have had a role to play in these changes.

Our findings suggest that PhilE patterns with its colonial parent, AmE, rather than with BrE, with a massive increase in *that* over the three decade period that raises its frequency to a similar level to that of AmE in the 1990s, and a rate of decline for *wh* that is such as to keep their 1960s and 1990s frequencies in tandem. Further confirmation of the close parallels between PhilE and AmE is to be found in the *wh* versus *that* ratios across the three macro-genres of press, learned writing and fiction. In all three genres the ratios for PhilE and AmE have come to be closely aligned in the 1990s, and clearly differentiated from the more conservative BrE, which consistently evidences stronger support for *wh* over *that*.

Given that relative *that* is known to be more commonly found in speech than in writing, its increasing popularity in written genres is undoubtedly being driven in part at least by the process of colloquialization (with prescriptive *which*-hunting a likely further factor). Another possible indication that colloquialism is a factor in relativization developments in PhilE is the finding that integrated relatives, known to be more commonly found in speech than writing, are on the rise while supplementary relatives are in (mild) decline. It is noteworthy that in Collins et al.'s (ms) recent study of modality in PhilE quasi-modals, also a colloquial feature, evidenced a rate of increase greater than that experienced in AmE, suggesting that it is in similar fashion to relative *that* attempting to bridge the frequency gap with its colonial parent.

Our findings provide evidence of continuing exonormative allegiance to AmE amongst PhilE writers and suggest that, at least on the basis of our findings for relativization, the case for PhilE having established itself in Phase 4 ('endonormative stabilization') of Schneider's (2007) evolutionary scale remains tenuous.

Notes

[1] The research for this study was supported by an Australian Research Council Discovery Grant. We also acknowledge the contribution of undergraduate students in the course "Language Research" (Term 1, Academic Year 2011-2012) taught by Ariane Macalinga Borlongan at De La Salle University for their help in compiling Phil-Brown.

[2] The time-consuming method used by Coronel (2010) was to manually search the whole of ICE-Phil, and subsequently run a concordance check of what Sigley (1997: 213) calls "likely environments for zero", such as all/something/nothing + personal pronoun. In Gut and Coronel (2012), similarly zero relative clauses were identified manually in all the corpus files used for the study.

[3] Because the compilation of Phil-Brown is still under way, no attempt has yet been made to match the precise word count per genre in Phil-Brown with the target word count of the Brown family. The strategy that the corpus compilers are employing at the current stage is to collect as much relevant data as possible. This explains the numerical differences between the word counts in the two columns, which makes normalization of the raw frequencies crucial.

4 In view of the stylistic heterogeneity of the general prose category, it is not surprising that the pioneering corpus-based Longman grammar (Biber et al. 1999) dispenses with it, concentrating instead on the same three written registers that we recognize in the present study: press, learned writing, and fiction.

5 See also Huddleston and Pullum (2002: 1034-1035).

6 Approximate frequencies can, however, be extrapolated from Biber et al. (1999: 610-611), albeit for BrE and AmE combined.

7 Coronel's normalized frequencies differ slightly from the present ones because her analysis was based on ICE-Phil in its entirety, whereas the frequency findings of the present study are derived from only half (204,000 words) of the written section of ICE-Phil.

8 The association between low frequency and vulnerability to attrition has been noted in the case of the modals by Leech et al. (2009: 73), where it is the low frequency items, particularly *shall*, *ought to* and *need*, that have recorded strong declines.

9 BrE is the only variety for which comparable diachronic data is available, as presented in Leech et al. (2009: 231-233).

References

Ball, C. (1996), 'A diachronic study of relative markers in spoken and written English', *World Englishes,* 17: 127-138.

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

Borlongan, A. (2009), 'A survey on language use, attitudes, and identity in relation to Philippine English among young generation Filipinos: an initial sample from a private university', *The Philippine ESL Journal*, 3: 74-107.

Borlongan, A. (2011), 'Relocating Philippine English in Schneider's Dynamic Model'. Paper presented at the *17th International Association of World Englishes Conference*, Melbourne, November 23-25.

Collins, P., A. Borlongan and X. Yao. (manuscript). 'Modality in Philippine English: a diachronic study'.

Coronel, L. (2010), *Relativization Strategies in Philippine English*. Unpublished Master's thesis, University of Augsburg.

Fowler, H. (1965[1926]), *A Dictionary of Modern English usage*. Oxford: OUP.

Gonzales, A. (1997), 'Philippine English: a variety in search of legitimation', in: E. Schneider (ed.) *Englishes around the World. Studies in Honour of Manfred Gorlach*. Amsterdam: Benjamins. 205-212.

Gonzales, A. (2004), 'The social dimension of Philippine English'. *World Englishes,* 23: 7-16.

Gut, U. and L. Coronel (2012), 'Relatives worldwide', in M. Hundt and U. Gut (eds.) *Mapping unity and Diversity in New Englishes*. Amsterdam: Benjamins, 205-241.

Guy, G. and R. Bailey (1995), 'On the choice of relative pronouns in English', *American Speech*, 70: 147-162.

Hoffman, T. (2005), 'Variable vs. categorical effects: preposition pied piping and stranding in British English relative clauses', *Journal of English Linguistics*, 33: 257-297.

Huddleston, R. and Pullum, G. (2002), *The Cambridge Grammar of the English Language.* Cambridge: CUP.

Kikai, A., M. Schleppegrell and S. Tagliamonte. (1987), 'The influence of syntactic position on relativisation strategies', in: K. Denning, S. Inkelas, F. McNair-Cox, and J. Rickford (eds.) *Variation in Language: NWAVE-XV at Stanford.* Stanford: Dept of Linguistics, Stanford University. 266-277.

Leech, G., M. Hundt, C. Mair and N. Smith. (2009), *Change in Contemporary English: a Grammatical Study.* Cambridge: CUP.

Quirk, R. (1957), 'Relative clauses in educated spoken English', *English Studies*, 38: 97-109.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Schneider, E. (2007), *Postcolonial English: Varieties of English around the World*. Cambridge: CUP.

Sigley, R. (1997), 'The influence of formality and channel on relative pronoun choice in New Zealand English', *English Language and Linguistics*, 1: 207-232.

Strunk, W. Jr. and E. B. White (1979 [1959]), *The Elements of Style*. New York: Longman.

Szmrecsanyi, B. (2012), '*That*, *which*, zero: prescriptivist influences on English relativizer usage'. Paper presented at the Freiburg Institute for Advanced Studies, Albert-Ludwigs-Universität Freiburg, May 9.

Tottie, G. (1997), 'Overseas relatives: British-American differences in relative marker usage', in: J. Aarts, I. de Monnink and H. Wekker (eds.) *Studies in English Language and Teaching. In Honour of Flor Aarts*. Amsterdam: Rodopi. 153-165.

The progressive in South Asian and Southeast Asian varieties of English – mapping areal homogeneity and heterogeneity

Marco Schilk and Marc Hammel***

*University of Hildesheim
**Justus Liebig University Giessen

Abstract

In the present study we use the South Asian and Southeast Asian components of the International Corpus of English *(ICE) as well as a larger set of web-derived newspaper corpora in order to account for similarities and differences in the use of the progressive in eight non-native varieties of English (Singapore, Hong Kong, Indian, Sri Lankan, Bangladeshi, Maldivian, Nepali and Pakistani English). The analysis is two-fold: First, for the varieties under scrutiny, we provide a quantitative overview of the different use of progressive aspect marking according to tense and voice. Second, we apply a cluster analysis in order to determine respective variety-based clusters so that a comparison of those verbs of which progressive aspect marking is said to differ most significantly will be feasible. In a third step we present examples of some of the most influential verbs that are responsible for the differences in use across the corpora.*

1. Introduction

Innovative use of progressive aspect has been identified as a characteristic feature of many New Englishes across the world (Platt, Weber and Ho 1984, Williams 1987, van Rooy 2006, Hundt and Vogel 2011). Firstly, there seem to be purely quantitative differences in the use of progressive aspect, with a general tendency in non-native varieties of English for a higher overall frequency of use, compared with British English (Schmied 1994, Rogers 2002). This may be partially due to the extension of progressive aspect marking to stative and habitual contexts, typically with verbs like *have* or *know*, where the use of the progressive is generally not admissible in native varieties (cf. e.g. Platt, Weber and Ho 1984: 72ff). Although this use seems to be typical of many outer-circle varieties, some observable differences between them have also emerged (e.g. progressive marking with the particle *still* in Singapore English (Wee 2004)), so that the degree of homogeneity between these varieties is yet relatively unclear.

Almost all linguistic features in regional varieties of Englishes around the world are shared with other varieties. Thus, it has been reported that deviations from an idealized global norm are relatively peripheral albeit pervasive in all varieties of English. In this context, especially quantitative approaches to corpus linguistics have contributed to the detection of these peripheral differences in use across different varieties of English. Accordingly, Schneider (2007: 87) writes:

> These are stable and noteworthy results, and it is worth pointing out that they operate way below the level of linguistic awareness: without quantitative methodology no observer would have expected such differences to exist (2007: 87)

Although differences in use are expected to exist below the overt level of linguistic awareness, we thus can identify regional morphosyntactic variation in use in terms of overall frequency as well as stylistic preferences. These distinctions in globally common patterns, such as the use of the progressive, are especially well-suited to test claims and hypotheses about dynamic and on-going language change and concerning different varieties of Englishes worldwide. Especially in the use of progressive marking and form-meaning pairings of the progressive, variation that seems to be unobtrusive at first glance may well be influential for the development of specific and distinctive regional uses.

The main focus of this paper is on variation in the use of the progressive across South Asian and Southeast Asian varieties of English. The results are compared against British English as their historical input variety. Variability in this particular area of grammar has been reported for inner-circle varieties (Mair and Hundt 1995, Smith 2002, Das 2010) and outer-circle varieties (Collins 2008, Sharma 2009, Filppula et al. 2009, Collins and Yao 2012).

One important factor for variation in inner-circle varieties seems to be an extension of the construction to future meaning (cf. Nesselhauf 2007). In the same vein, Smith (2005) discusses some features that seem to be strongly associated with the increasing use of the progressive in British English: grammaticalization (i.e. gradual generalising of the progressive from its original meaning to more abstract meanings) and the influence of American English on British English and other varieties (Filppula et al. 2009: 256). Römer (2005: 173), suggests that the reason for a steady increase in the use of the progressive can be put down to “inadequate descriptions of language phenomena in teaching materials”, whereas Platt et al. (1984: 73) assume that the wider use of the progressive is partly due to overteaching of this construction. In SLA research, the grammatically correct use of the English progressive has often been regarded as one of the most difficult things to learn. Swan and Smith (2001: IX) list the progressive as one of most problematic areas for almost all learner groups. Furthermore, Housen (2002) states that the extension of the use of the progressive is rather contingent upon the level of proficiency of an individual speaker. With respect to varieties of English, a lectal cline of speakers may thus be assumed where less proficient speakers of an ESL variety share linguistic characteristics with intermediate and advanced EFL learners.

These considerations aside, there is common agreement that the use of the progressive considerably differs functionally between L2 and L1 varieties of English. Notwithstanding, as noted by Quirk et al. (1985: 198) and Biber et al. (1999: 461-463), since the progressive seems to be more frequent in speech than in writing, studies based solely on written corpora, such as the LOB/Brown family, will remain somewhat inconclusive. Outer-circle varieties – apart from using a considerably higher frequency of the progressive – seem to extend the construction to contexts of verbs that are mainly stative in meaning (Platt, Weber and Ho 1984: 72-73, Williams 1987: 172-173, Filppula et al. 2009). Schmied (1994: 223) shows that there are more instances of progressive marking in the Indian Kolhapur corpus compared to the British LOB and American Brown corpora. In the same vein, Rogers (2002: 193) writes that "I noticed that the progressive form seemed to occur more frequently than one might have expected". Meshtrie (2004: 1134) reports that the use of the progressive with stative verbs is so widespread in the L2 varieties of English in Africa and Asia that it may well be considered an almost universal feature.

Various attempts have been made in order to explain the varied use of the progressive across different varieties of English. Some scholars, for example, put down the differences in use to substrate influences of the ESL speakers' respective L1s. Thus, L2 speakers of English are said to use this pattern in a fashion that deviates from ENL varieties either because it is missing from their mother tongue as a grammatical construction or because there are usage differences between English and the respective speakers' L1s. Such differences in use may, for example, explain the extension of progressive marking to stative verbs or habitual processes.

> Although explanatory attempts concerning regional differences are not unanimous, there seems to be agreement that differences may be explained by colloquialization, a process that is responsible for narrowing the gap between the norms of written and spoken English (Mair 1998: 148).

Some earlier studies document the general diachronic tendency of a steady increase of frequency of the progressive aspect since Late Modern English (e.g. Elsness 1994, Smitterberg 2005, Nesselhauf 2007). Moreover, Mair and Leech (2006) report that the progressive is still on the rise and that it is becoming established in the "few remaining niches of the verbal paradigm which it was not current until the 20th century" (Mair and Leech 2006: 323), i.e. it is now common to encounter e.g. combinations of the progressive with modals and the passive voice.

This phenomenon of gradual increase in the use of the progressive may also be explained by what Ranta (2006: 114) calls 'attention catchingness'. This statement is corroborated by Haspelmath (1999: 1057) who writes that speakers strive to use language in an innovative or unexpected way in order to be noticed. Furthermore, the worldwide McDonald's advertisement of 2005 *I'm loving it*, may also be viewed as an influential reinforcement of the increasing use of the progressive (cf. Mesthrie and Bhatt 2008: 67).

Thus, the phenomenon of the extended use of the progressive is probably not simply a substrate dependent feature of general interference from an individual L2 speaker's L1, since similar use is attested in the speech of L2 speakers from many typologically different mother tongues (Ranta 2006). A possible assumption is the attractiveness of the progressive residing in the communicative value in interaction. Jenkins (2000: 160) observes: “There really is no justification for doggedly persisting in referring to an item as an error if the vast majority of the world’s L2 speakers produce and understand it”.

What prima facie seems to be a ‘universal’ overuse of the progressive can probably be carved out very differently in individual varieties of English. Nevertheless, Sharma (2009) concludes that although Singapore English and Indian English both overuse the progressive when compared to other varieties, these usage patterns can only be fully understood in light of the respective substrate languages involved (Hindi for Indian English and Tamil for Singapore English).

While the South Asian and Southeast Asian varieties of English under scrutiny belong to the category of outer-circle varieties in the Kachruvian sense (cf. Kachru 1992), it thus seems useful to characterize them not only in terms of this monolithical classification, but also by the types of linguistic innovations that are the result of localized functions the language is supposed to carry out, and in terms of the adaptation of communicative strategies and transfer from local languages. This is expected to go hand in hand with the progressive taking on new functions that are unknown to inner-circle varieties of English (cf. van Rooy: 2006: 39). Therefore, we expect to come across quantitative as well as qualitative differences with respect to the use of the progressive across the varieties of Englishes to be examined.

The present analysis thus tries to extend the scope of previous studies by analysing eight Asian English varieties and exploring possible ramifications. It includes both written and spoken data and attempts to shed light on presumable accountability via an examination of various variables. The following questions will be addressed:

a) Do South Asian and Southeast Asian Englishes use more progressives for specific verbs than British English? (Do frequency-based progressive profiles of different varieties exhibit areal homogeneity?)
b) Do South Asian and Southeast Asian Englishes provide examples of stative verbs in the progressive and do they use them to the same extent? Will we find aspects of usage in which Asian Englishes are different from British English? (Which functional differences can be shown in the use of the progressive across varieties of English?)
c) Are there genre-specific differences?
d) Are there any other new forms that have been established with the combination of the progressive?

2. An overview of typical uses of progressive aspect marking

The progressive aspect in the English language is prototypically expressed by a combination of the auxiliary *BE* with the present participle *-ing*. The progressive aspect focuses on the situation as being in progress at a particular time, commonly limited in duration, and not necessarily completed at the time of speaking (cf. Quirk et al. 1985: 197). Consider examples (1-2):

(1) I read the newspaper yesterday morning.
(2) I was reading the newspaper yesterday morning.

In (1) the speaker emphasizes the completion of the action, whereas in (2) the temporary nature of this utterance shifts into focus. Due to semantic differences, the progressive affects various verb categories in different ways. Stative verbs do not occur in the progressive, since there is no conception of progression in states of affairs (cf. Quirk et al. 1985: 198). That is why the following examples (3) and (4) are semantically incompatible with the progressive:

(3) I am liking your brother.
(4) I am being a girl.

However, when verbs that are ordinarily stative occur in the progressive, they adopt dynamic meanings, so that the construction adds to the meaning of the verb. They may show a type of behaviour of limited duration such as in (5):

(5) He was being crazy.

Furthermore, when the progressive is used with verbs expressing emotions or attitudes or locatives, which are ordinarily stative, they indicate tentativeness or temporariness as in (6-7):

(6) I am hoping to take my exam next Friday. (tentative)
(7) I was staying at Ian's. (not a permanent residency)

When the progressive occurs with verbs denoting states a temporary or dynamic interpretation must be conceivable (cf. Quirk et al. 1985: 198-206). With durative actions, the progressive is interpreted as conveying the incompleteness of activity, process or change as illustrated in examples (8-10):

(8) Tom is studying.
(9) The weather is getting colder.
(10) Bridgit is writing a term paper.

Besides these semantic factors, there are other constraints that are said to impair the distribution of the progressive aspect in English. In Biber et al. (1999: 461-462) the perfect progressive is reported to be used very scarcely across all registers (only in 0.5 % of all verb phrases).

Secondly, highly complex constructions, above all the perfect progressive passive (as in 11), seem to be avoided because it is "felt to be awkward" (Quirk et al. 1985: 213):

(11) The road has been being repaired for months.

Apart from avoiding particular highly complex constructions, there also seems to be some variation with regard to dialect: Biber et al. (1999: 462) state that speakers of American English more frequently use the progressive than speakers of British English, especially in conversation.

This brief summary of the use of the progressive in English shows that differences in form and function are to be expected in the forthcoming analysis, since the progressive construction is still undergoing change in international varieties of English.[1]

3. Database and methodology

For our analysis we used two different datasets, a large-scale newspaper corpus of South Asian and Southeast Asian varieties of English and a dataset consisting mainly of the corresponding ICE-corpora. The newspaper corpus consists of data from 9 varieties of English with a total of about 22 million words. The ICE-based dataset is much smaller, since there are no corresponding ICE-corpora for all varieties covered by the newspaper data. For British English, we used the newspaper texts contained in the British National Corpus (BNC). Table 1 presents an overview of the data used for the present study.

Table 1. South Asian and Southeast Asian Corpora

Newspaper Corpora (SAVE & Southeast Asian)		ICE	approx. size
Bangladesh	Daily Star	n.a.	1.5m
	New Age	n.a.	1.5m
India	The Statesman	ICE-India	1.5m | 1m
	The Times of India	n.a.	1.5m
Maldives	Dhivehi Observer	n.a.	1.5m
	Minivan News	n.a.	1.5m
Nepal	Nepali Times	n.a.	1.5m
	The Himalayan Times	n.a.	1.5m
Pakistan	Daily Times	n.a.	1.5m
	Dawn	n.a.	1.5m
Sri Lanka	Daily Mirror	ICE-SL$_{W200}$	1.5m | 0.4m
	Daily News	n.a.	1.5m
Hong Kong	South China Morning Post (SCMP)	ICE-HK	1.5m | 1m
Singapore	Straits Times (ST)	ICE-Sin	1.5m | 1m
Great Britain	BNC$_{\text{NEWS}}$	ICE-GB	1.5m | 1m

The South Asian newspaper corpora are web-derived newspaper corpora that contain three million words per variety; the data was generated from the online archives of two different newspapers in each case. This data is complemented with data from two Southeast Asian newspapers containing 1.5 million words of data each.

As the *International Corpus of English* does not contain data for all represented South Asian varieties, this data is restricted to the Indian component of ICE and the written part of ICE-SL. Since differences in the use of the progressive may be more marked in spoken language that is not contained in the newspaper corpora, we also added a number of other varieties to the ICE data. These are three native varieties (Canadian English, Irish English and New Zealand English) and a further ESL variety (ICE-PHI). Since for these varieties no corresponding newspaper corpora exist in our setting, they are not shown in Table 1.

In order to compare progressive aspect marking across the varieties contained in the two corpus environments, we added part-of-speech tagging to those corpora that did not already exist in a POS-tagged format. Since the South Asian Varieties of English newspaper corpora (SAVE) are annotated with the CLAWS7 tagset, we decided to use the same tagset for all corpora under scrutiny.[2]

Based on this tagset we used a regular-expression-based search algorithm to create lists of progressive uses in the different varieties. Since the objective of this procedure is to create lists that are quantitatively comparable across all varieties contained in the study, we decided on a high precision approach that comes at the cost of recall: our regular expression only identifies those uses of any given verb as a progressive use if the syntactic representation of the progressive follows a relatively simple particular pattern. This pattern is defined as a verb containing the continuous inflectional morpheme *-ing* that is directly preceded by a form of the auxiliary BE.[3] While this approach only identifies progressive uses of verbs in the corpora, it does not identify all of those uses as instances where, for instance, a modifier is used between the auxiliary and the main verb are not identified. As this classification is only used to identify the verbs that display the highest variation in progressive marking across the varieties but is not used for the qualitative interpretation of the variation, we opted for precision over recall. The specific verbs identified to vary between the different regional varieties are in a second step analysed in more detail. During this closer analysis the occurrences that are not identified by the automatic procedure are also taken into account.

The automatically generated list thus contains all instances of verbs with explicit progressive marking that are identified by the regex-procedure and can, therefore, be seen as a progressive profile of a specific variety. To compare the use of the progressive across varieties, these profiles were used as the basis for a statistical cluster analysis. Based on a Ward's method distance matrix we identified which varieties are significantly similar to each other, i.e. formed variety-based clusters.[4] The cluster analysis was performed with the R package {pvclust}. The hierarchical clustering is based on Ward's sum of squares method (Ward 1963), multiscale bootstrap resampling with 1000 replications was used to test for statistical significance (approximately unbiased probability values (AU)) (cf. Suzuki and Shimidora 2006). The first results of our analysis will be shown in the form of variety-based clusters. Varieties within each cluster are highly heterogeneous in their use of the progressive across the whole spectrum of different verbs, while variation can be identified across the different clusters.

In a second step we compared the different variety-based clusters to each other. During this phase of the procedure, we identified those verbs that display the strongest amount of variation between the respective clusters. By using the squared sums of the residuals of the distributions and comparing the partial variation of influential verbs to the total variation within the distribution, it was possible to identify the verbs that are used differently and quantify this variation as a percentage of the total variation between the clusters.

In a third step we chose candidates for a qualitative analysis from the high-variation verbs. The choice of those candidates is partially motivated by the amount of variation these verbs display and partially on a more intuitive level. For example, if we found so-called stative verbs amongst those that display a high amount of variation, we subjected these to a further qualitative analysis. During this analysis we only take a number of specific verbs into consideration, but in contrast to the automatic procedure we consider all instances of these specific verbs – without the regex-based constrictions – to shed light on the possible motivations for differing use of progressive aspect between the varieties.

4. Results

4.1 Progressive use in the newspaper corpora

Our first dataset includes the nine newspaper corpora, i.e. the seven South Asian corpora included in the SAVE corpus, the two Southeast Asian corpora and the newspaper section from the BNC. After the identification of all progressive verbal uses in these nine corpora, the corpora that display significant similarity in progressive aspect marking of all the included verbs were grouped together. The resulting groups of corpora are visualized as a cluster dendrogram in Figure 1.

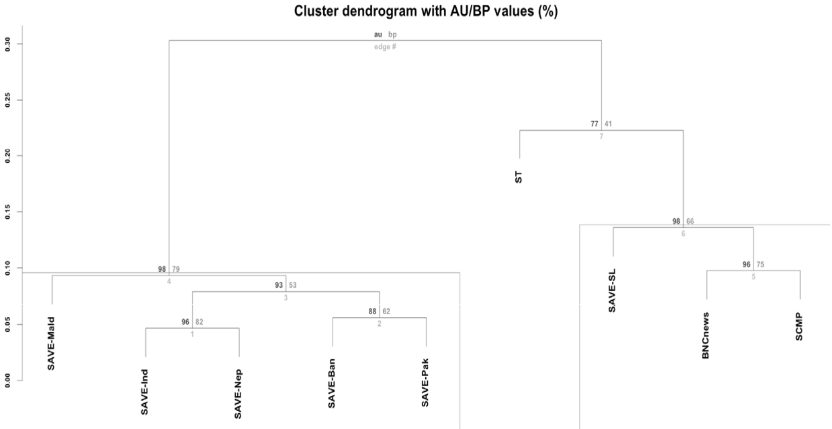


Figure 1.

Figure 1 shows the result of the cluster analysis performed on the verb-profiles of the newspaper data. In the nine corpora under scrutiny our analysis identified two homogeneous groups, i.e. groups of corpora in which the use of the progressive is significantly similar. On the left-hand side you can see that most of the South Asian Varieties are grouped in one cluster (Maldivian English, Indian English, Nepali English, Bangladeshi English, Pakistani English), the exception being Sri Lankan English. Sri Lankan English is grouped together with Hong Kong English and British English. Singapore English is not part of either cluster.

This cluster analysis shows that, firstly, most South Asian varieties behave similarly in terms of progressive aspect marking, so that at least on this level of description the cover term ‘South Asian Englishes’ seems to be justified on linguistic as well as regional grounds for the included varieties.

Interestingly, Sri Lankan English is not part of the South Asian cluster, a result that is in line with some earlier studies where Sri Lankan English was shown to be closer to British English than other South Asian varieties, such as Indian English (cf. Schilk et al. 2012). A further noteworthy point is that Hong Kong English is grouped with British English, while Singapore English is not included in any of the clusters. From a developmental perspective this may strengthen earlier assumptions that while Hong Kong English is relatively close to British English and perhaps fossilized in stage two of Schneider’s (2007) model of World Englishes, Singapore English displays unique characteristics that indicate further varietal development and ‘endonormative stabilization’.

On the basis of the cluster analysis of the verb-profiles of the newspaper data, we collapsed our verbal profiles according to the two significant clusters in order to identify which verbs are mainly responsible for the variation between the groups by looking at the residuals of the distribution. As can be seen from the following table, from the 2,922 verbs that are covered in total, only 25 verbs account for more than 23% of the total variation between the clusters.

Table 2. Comparison of variation between SAVE and BNC/SL/SCMP data

Verb (.vvg)	BNC/SL/HK	SAVE (no SL)	% of Variation	Total % of Variation
hoping	11.67	- 11.24	3.56	3.56
looking	10.24	-9.87	2.74	6.30
trying	- 8.99	8.66	2.11	8.41
appealing	8.10	-7.80	1.72	10.13
facing	- 7.59	7.31	1.51	11.63
seeking	6.47	- 6.23	1.09	12.73
demanding	- 6.19	5.96	1.00	13.73
playing	5.56	- 5.36	0.81	14.53
undergoing	- 5.37	5.18	0.75	15.29
protesting	- 5.30	5.10	0.73	16.02
beginning	5.17	- 4.98	0.70	16.72
recovering	5.07	- 4.88	0.67	17.39
absconding	- 5.03	4.85	0.66	18.05
offering	5.00	- 4.81	0.65	18.71
working	- 4.88	4.70	0.62	19.33
drinking	4.49	- 4.33	0.53	19.86
aiming	4.45	- 4.29	0.52	20.37
passing	- 4.29	4.13	0.48	20.85
heading	4.21	- 4.05	0.46	21.32
contesting	- 4.07	3.92	0.43	21.75
increasing	- 3.94	3.80	0.41	22.15
participating	- 3.93	3.79	0.40	22.56
using	- 3.76	3.62	0.37	22.93
wearing	3.71	- 3.58	0.36	23.29

Table 2 displays the main verbs that are responsible for the variation between the two significant clusters in decreasing order of variational difference. The figures in the second and third column are the residuals of the distribution of the data, positive and negative values indicate ‘overuse’ and ‘underuse’ with respect to the other cluster. The squares of those residuals can then be compared against the sum of squared residuals of the total distribution, thus identifying the percentage of total variation each verb is responsible for.

In the following we take a closer look at two of these verbs, namely *try* and *work*, for which we have discovered some examples that may account for some of the qualitative differences in use. For the purposes of this discussion, we confined ourselves to the examination of the cluster of South Asian Varieties. In order to interpret the quantitative findings in terms of quality we looked for concordances that differ from the prototypical instances described in Section 2.

For *trying* we have detected several examples of the progressive without the copula, as examples (12-13) from the Maldivian "Minivan News" illustrate:

(12) [...] so right now he trying to find an excuse for reform (Minivan News – Maldives)

(13) Aree had previously been sent away after he trying to obstruct the arrest (Minivan News – Maldives)

In (12), the speaker is currently looking for an excuse that explains the reform. From the perspective of tense use here, the progressive is applied prototypically since we are dealing with a description of a situation that is in progress. The only essential part of the progressive construction that is lacking is the copula form of the verb BE.

Since the omission of the copula in combination with the progressive construction is likely to be due to informal speech style among educated speakers of English in spoken language, it can be assumed that this pattern is sometimes transferred to the written medium.

In (13), the protagonist Aree had been sent away after he tried to obstruct the arrest. In this example, in addition to the lack of the copula, we also find a possible deviation in the sequence of tenses. However, due to the missing copula the intended tense is not entirely identifiable. One would either expect a simple form instead of a progressive construction in this case, or tense marking in the realization of the copula.

An explanation for the absence of the appropriate form of *be* may lie in the different acquisitional processes of L1 and L2 speakers. In Brown (1973) the present progressive was said to be the first grammatical morpheme to be acquired in L1-acquisition and in L2-acquisition, the progressive marker *-ing* was also among the first morphemes acquired as discussed in studies such as Cook (1993) or Gass and Selinker (2001). Thus, the example *he trying* could be classified as appropriate use of the progressive (e.g. in Gass and Selinker 2001).

A further case of non-standard use of the progressive is the subsequent combination of two progressive forms one after another, as shown in examples (14-15):

(14) At the moment it is said that the Indian embassy in Kathmandu has being trying to get the Nepalis, wounded by the negotiations to meet again (New Times – Nepal)

(15) We are making trying to take Indian cinema to a global market in a systematic way like they do in business (Times of India – India)

In (14) the author used a combination of two present participles one after another. Since the present perfect progressive is often used to focus on the duration of an action, it is possible that the author uses the progressive *being* instead of the past participle *been* in this example. The use of *being* instead of *be* may also be motivated by a possible (partial) homophony of *being* and *been* in the author's dialect. However, as author identification is notoriously difficult in the case of newspaper corpora, it is not possible to follow up on this possibility.

In (15) we find a combination of two present participles in active voice. Such a combination into a 'stacked progressive' is in this case clearly a marked option. A possible explanation for this use is an overextension of the progressive form *trying* to serve as a substitute for a noun phrase with the meaning of *attempts* or *efforts*. This use of a stacked progressive is very rare, so we clearly treat this interpretation conservatively in terms of typicality. As, however, extensions of form-meaning mappings in the form of structural semantic analogies have been shown to exist in Indian English on other levels of the lexis-grammar continuum (Mukherjee and Hoffmann 2006), quasi-nominal extension of the progressive form of verbs may well be responsible for innovations such as this.

For the second sample verb *working*, we have also come across examples that may account for the variation in the use of the progressive across the clusters. Examples (16-17) show the combination of the present progressive with an adverb of time (*since*, *for*) that would require a present perfect in British English.

(16) WWF India is working in the Ladak region since November 1999 to help conserve the fragile altitude wetlands (The Statesman – India)

(17) The farmers, agricultural scientists, and policy makers are working hard for years together (Daily Star – Bangladesh)

In both examples (16) and (17) a situation that has started in the past and lasts until the present is expressed with the present progressive. Speakers of British English would prefer to use the present perfect progressive instead of the present progressive here. However, the use of the present progressive instead of the present perfect progressive is said to be among the major sources of variation with speakers of ESL-varieties (cf. Swan and Smith 2001).

What should be mentioned in this respect is that variation in tense use is not necessarily responsible for frequency differences in the use of the progressive as reported by previous studies such as Hundt and Vogel (2011). Hundt and Vogel (2011: 159) state that the extension of the progressive to contexts of perfective marking may be due to the fact that the past progressive is similarly used, namely as a marker of recent past. However, it should not go unnoticed that the main example they quote is slightly different: there is a stronger element of resultative meaning than in (16-17), where the focus seems to be on the continuation of a situation that began in the past.[5]

Since overuse of the progressive construction has often been reported to be majorly due to the use of stative verbs in the progressive, we further searched the corpora for the following stative verbs in their progressive form:

> admire, adore, **appear**, astonish, believe, belong, concern, consist, deserve, desire, despise, detest, dislike, doubt, envy, exist, fit, forget, guess, hate, have, hear, imagine, impress, include, involve, keep, know, lack, last, like, love, matter, mean, owe, own, please, possess, prefer, reach, realize, recognize, remember, resemble, satisfy, see, seem, sound, smell, stop, suppose, surprise, survive, suspect, understand, want, wish (cf. Quirk et al. 1985: 200-213)

It is surprising that out of these verbs only *appear* is among the top 100 verbs that differ in terms of frequency of use in the progressive (rank 78, 0.17% of total variation) across the different newspaper corpora. Several instances where *appear* is used with marked progressive marking are shown in (18-25):

(18) Abdulla Kamaluddheen, one of the members and a candidate for the majlis election on 31 December, is appearing regularly in the Maldives media in an attempt to boost his weak support base (Dhivehi Observer – Maldives)

(19) Advertisements have been appearing on print and electronic media for booking flats and shops in the project. (Dawn – Pakistan)

(20) Subhas Chakraborty is appearing more politically correct these days (Times of India – India)

(21) This North American snowy egret is reportedly appearing to be headed south for a warmer climate but by then it has attracted hundreds of bird watchers since it was first spotted in Britain last October (Times of India – India)

(22) Sohrowardi Shuvo is appearing in the HSC examination (New Age – Bangladesh)

(23) Now, according to the newspapers, Alzheimer is appearing in younger and younger people across Canada and the United States (Nepali Times – Nepal)

(24) The couple rent a home in the resort Joe is appearing for the summer (SCMP – Hong Kong)

(25) The Swinging Blues Jeans are appearing at the Mayfair Center Seaton Carew tonight (BNC – Great Britain)

Example (18) illustrates an overextension of the progressive construction into habitual contexts. Since *regularly* is an adverb that is used in order to express habituality, we may be dealing with a non-prototypical use of the progressive construction here. However, this interpretation is somewhat controversial since the use of the progressive in this example may also indicate the temporary duration of this event.

In the example from the Pakistani Dawn Corpus (19), the progressive construction again seems to have been stretched to non-prototypical contexts. By definition, the present perfect progressive is used when the speaker or writer wants to put emphasis on the duration of an action. Contrariwise, in this respect the present perfect progressive is used to convey a result, namely, that advertisements have appeared on print and electronic media. Speakers and writers of British English would probably prefer the present perfect simple in this case.

In example (20) from the Times of India newspaper corpus, the leader of the Indian Communist Party is described as being more politically correct than he usually is. This use of the progressive in this context seems to be marked inasmuch that a characteristic trait of a person does not commonly change and one would probably expect a simple form in this context.

In the second example from the Times of India newspaper corpus (21), the the progressive is similarly used. Since birds such as the snowy egret always head south during the winter time, the progressive construction in this case may also be regarded as another example of overextension of the progressive aspect into habitual contexts.

Besides these findings of uses of the progressive extended to the stative verb *appear* in the cluster of South Asian varieties (without Sri Lanka) of English, there are also some similarities with the BE – HKE – SLE cluster. The progressive construction seems to be commonly used in order to express future actions, as shown in (22-23). In both examples, the progressive is used to refer to future events. In the same vein, examples (24-25) from the Hong Kong-corpus and the British newspaper corpus neatly illustrate the extended use of the present progressive to refer to future actions.

As an interim result of the newspaper data it thus seems to be the case that differences in the use of the progressive are rarely due to the extension of use of progressive to so-called stative verbs, such as *appear*. Although we do find a number of examples where this is the case, the verbs that are mainly responsible for the variational differences do not fall into the category of stative verbs. However, we do find marked uses of the progressive with the two example verbs *trying* and *working*. In this case we mainly find an extension of the progressive to habitual contexts or the use of 'stacked' progressives, where the second progressive may substitute a semantically similar noun phrase.

4.2 Progressive use in the *International Corpus of English* (ICE)

The second part of our analysis is based on eight different ICE-corpora, four of which cover varieties that were also looked at in the newspaper dataset. Initially, we performed a cluster analysis on the ICE data in the same fashion as in Section 4.1.

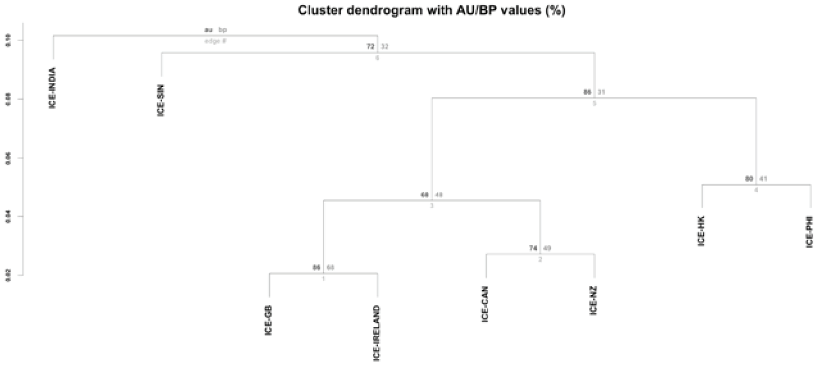


Figure 2. Progressive marking in ICE-corpora

The dendrogram in Figure 2 shows that ENL varieties of English seem to form one cluster, while HKE and PhilE form a second cluster, whereas SinE and IndE are not part of these clusters. While this makes sense intuitively and is in line with some of our earlier results, resampling shows that none of these clusters are statistically significant. We assume that this is partially due to the fact that the ICE corpora contain a variety of different text-types from both the spoken and the written mode, which will lead to some variation in the use of the progressive within the ICE corpora.

By stratifying the data according to text-types we can see that the clusters are mainly medium and text-type based and that there is no cluster that would contain, say, only one variety (cf. Figure 3). Therefore, we will look more specifically at the largest spoken text-type, namely dialogues, because this text-type is arguably the one that differs most strongly from the newspaper data analysed in the first part of the study. When it comes to dialogues, we can see that the native varieties GB, NZ, and IRE build one significant cluster, whereas India and Hong Kong seem to be quite different from all other varieties (cf. Figure 4).

By looking at the residuals of the distribution we can again identify those verbs that display the largest variation between the varieties (cf. Table 3). Although *studying* is at the top of this list owing to an ‘overuse’ in Hong Kong, we assume this difference to be largely topic-based, i.e. ICE-HK containing many student dialogues for reasons of corpus-building economy. The more interesting verb in this distribution is *know*, as it has often been assumed that the use of so-called stative verbs is one of the fields where variation between native and second-language varieties occurs.

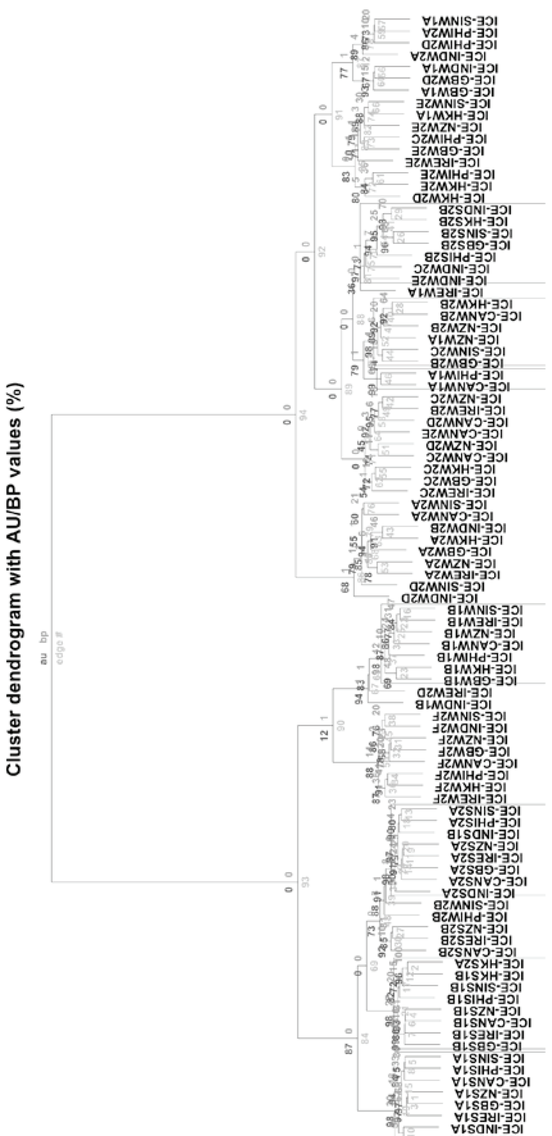


Figure 3. Progressive marking in ICE by variety and text category

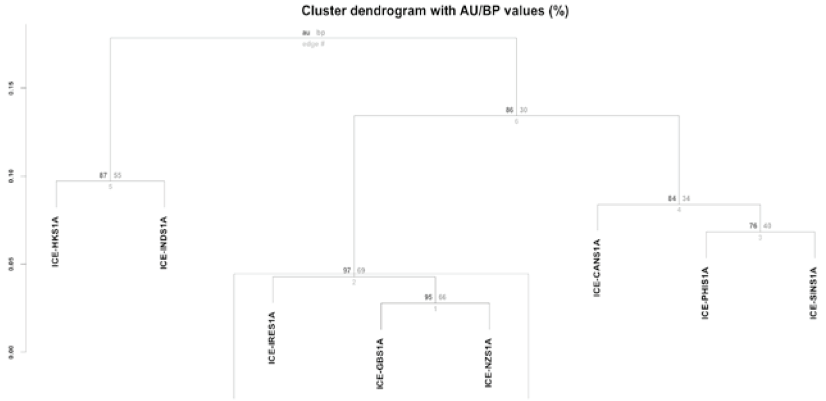


Figure 4. Progressive marking in ICE-S1A (dialogues)

Table 3. Comparison of variation in ICE – spoken dialogue (S1A)

Verb (.vvg)	ICE-CAN	ICE-GB	ICE-HK	ICE-IND	ICE-IRE	ICE-NZ	ICE-PHI	ICE-SIN	% of Variation
studying	-2.96	-2.76	12.79	2.09	-2.60	-3.77	1.07	-3.63	2.79
planning	-1.91	-2.64	0.89	1.83	-2.57	-1.03	8.11	-2.34	1.19
saying	-1.49	2.29	-3.14	-2.48	5.80	-0.05	-2.10	0.44	0.78
teaching	-1.21	-2.47	3.97	3.31	-3.29	-2.67	2.26	0.57	0.73
thinking	5.22	-1.55	-1.17	-2.14	-2.03	-1.40	1.61	1.89	0.61
suffering	-1.31	-0.60	0.08	6.28	-1.51	-1.40	0.14	-1.35	0.60
joining	-1.07	-1.10	0.23	6.37	-1.23	-1.14	-1.10	-0.19	0.60
knowing	-0.96	-0.98	0.00	6.32	-1.10	-1.02	-0.98	-0.98	0.58
sitting	2.33	-1.15	-1.09	-1.93	4.67	0.74	-2.30	-1.85	0.54
enjoying	-1.88	-0.38	-1.47	2.86	-1.71	-1.52	4.28	0.13	0.48
getting	1.70	0.88	-1.34	0.56	0.95	3.33	-2.89	-3.36	0.47
learning	-0.27	-1.74	4.90	-1.22	-1.66	1.65	-1.74	0.07	0.47
dating	-0.76	-0.78	-0.79	-0.76	-0.87	-0.81	5.67	-0.78	0.46
working	-0.43	0.21	1.62	4.32	-0.07	-1,83	-2.46	-1.23	0.41
cooking	-0.32	-1.20	-1.23	-0.33	-0.61	5.14	-1.20	-0.37	0.40
going	-2.18	1.22	-1.07	-1.85	3.45	1.92	-1.69	-0.38	0.37
forcing	-0.68	-0.69	4.93	-0.68	-0.78	-0.72	-0.69	-0.70	0.35
chatting	-0.76	-0.78	-0.79	-0.76	4.86	-0.81	-0.78	-0.78	0.35
graduating	-1.07	-0.18	4.23	-1.07	-1.23	-1.14	1.64	-1.10	0.34
looking	0.56	0.62	0.36	-3.97	-1.11	-0.21	1.14	2.69	0.34
staying	-0.75	-1.49	0.46	4.47	-0.77	-1.42	-0.59	0.30	0.33
coming	-1.24	1.52	-2.23	1.19	1.96	1.50	-3.06	0.11	0.33
applying	-1.35	4.38	-0.71	-0.62	-0.92	-0.06	-1.39	0.77	0.32

For our analysis of the sample verbs *studying* and *knowing*, we included all forms of BE + V-*ing*, allowing for up to three elements in between. A first result is that, apart from the topical differences mentioned above, we also find an extended use for the progressive with both speakers of HKE and IndE. Speakers of IndE and

HKE seem to use *studying* not only for the process of studying a specific subject but also for the state of being a student – especially in combination with location.

Table 4. *Studying* + NP/PP collocate in ICE spoken dialogue (S1A)

studying	**GB**	**IRE**	**NZ**	**CAN**	**PHI**	**SIN**	**HK**	**India**
total	6	2	-	3	6	-	14	9
subject	2	1	-	-	2	-	3	-
purpose	1	-	-	1	-	-	2	-
location	-	-	-	-	-	-	2	4
other	3	1	-	2	4	-	7	5

We have classified the different contextual uses of *studying* into four different categories: studying a subject, studying to become X or Y, studying at a location, and ʽother’. Examples (26-27) are typical examples from the British component of ICE:

(26) Well, when I was uh studying for my degree I was doing a joint a joint course part of which was visual arts or fine art and part of which was dance (ICE-GB: S1A-004)
(27) And I think one of the things that I felt when I was studying dance was I very much enjoyed the work that I was involved in […] (ICE-GB: S1A-001)

Speakers of British English tend to use the progressive form of *studying* along with the purpose what they are striving for as in (26) where the speakers claim to be studying for a degree. Moreover, it seems to be common to use the progressive aspect along with the object that people are currently studying as in (27), where the speaker talks about studying dance.

In the Irish data we also find a combination of *studying* with the actual object to be studied, as in (28). The example from ICE-Canada (29) is quite different from the other ones from the inner-circle corpora. Since this utterance contains a time adverbial such as *two or three days*, the speaker is supposed to stress the duration of the studying process by using the progressive aspect. Thus, this example neither fits the category of studying for a purpose nor the category of studying a certain object. Note, however, that in our cluster analysis Canada was not sorted among the other ENL varieties but together with ICE-PHI and ICE-Sin.

(28) Have you been studying several words […] (ICE-IRE: S1A-016)
(29) He’d been studying two or three days in a row (ICE-CAN: S1A-095)

Apart from studying a subject or for a certain purpose it seems to be typical of both Hong Kong English and Indian English to also refer to a location where somebody is studying. In (30), the HKE speaker explicitly refers to the time of having been a student in Japan.

(30) When I was studying in Japan […] (ICE-HK: S1A-059)
(31) Yes he is studying in the uh Chinese University (ICE-HK: S1A-017)

The same holds true for example (31), where the speaker puts emphasis on the location where he is studying, namely at the Chinese University. Similarly, this does not seem to be a proper combination of the progressive aspect in native varieties of English. Once again, one would rather expect utterances such as *when I used to be a student, I attended a Chinese University*. As illustrated in (32-33), the examples found in the Indian component are fairly similar to the ones yielded from its Hong Kong equivalent. In both examples, the speaker uses the progressive aspect of *studying* along with a location.

(32) My son is studying in a local school in uh fourth standard (ICE-India: S1A-029)
(33) My son is studying in Padmasheshadri the school in Madras I see (ICE-India: S1A-029)

Thus, the use of the progressive form of *study* in collocation with noun phrases that denote places seems to be relatively typical of both Indian English and Hong Kong English. Here the progressive is not used to refer to the actual process of studying but is rather denoting the state of being a student in a particular place. Consequently, although *study* is clearly not a stative verb, in these contexts the progressive is used with a stative meaning, while such instances are not attested for the cluster containing BrE, IRE-E and NZE.

The second sample verb that we analyse in more detail for the ICE data is *know*, which can be seen as a stative verb. Here we mainly find instances in the spoken component of ICE-India (twenty-two in the spoken data and three in the written data) and a few instances in ICE-HK (two instances in the spoken data).

By looking at the dialogue data we can find a number of examples of the stative verb *know* used with progressive inflection (examples (34-35) from ICE-HK, examples (37-39) from ICE-India).

(34) […] as if a person is **knowing** that he's going to die (ICE-HK: S1A-013)
(35) Did you find that one are proper not **knowing** […] (ICE-HK: S1A-084)
(36) "You must be **knowing** about the latest man in Sash's life?"[…] (ICE-SL: W1B-014)
(37) I probably may not be **knowing** […] (ICE-India: S1A-073)
(38) You must be **knowing** this all history […] (ICE-India: S1A-066)
(39) And these women, uh village women are not **knowing** anything about […] (ICE-India: S1A-088)

These findings for Indian English confirm Filppula et al. (2009: 253) who state that Indian English exhibits the most varied use of the progressive with stative verbs. Furthermore, Kachru (1983: 78) considers Indian English usage to be contact-induced and to represent an extension of the Hindi-Urdu pattern with stative verbs to Indian English. Moreover, the findings confirm Rogers (2002: 192) who finds instances of non-stative *knowing* in the Kolhapur Corpus of Indian English: "They were all *knowing* one another very well" (Rogers 2002: 192).

Comparing our findings for the ICE dataset to those for the newspaper corpora reveals several noteworthy points. Firstly, with data that is not limited to a specific genre, our cluster analysis does not seem to perform well enough. Although the clusters built from the complete ICE-corpora can serve as a useful heuristic, they lack significance when using multiscale bootstrap resampling. This is not entirely surprising, since differences between spoken and written language and differences between single genres tend to obscure regional preferences for specific linguistic realizations (Schilk, forthcoming). Limiting our approach to a single text type, however, shows that the differentiation between speech and writing in the two parts of the analysis helps to identify linguistic overlap and variation. An interesting case in this regard is HKE that was grouped with BrE in the analysis of the newspaper data but together with an ESL-variety (IndE) in the ICE-based analysis of conversations. This may be an indication that the use of the progressive is more nativized and standardized in South Asian Varieties of English where variety-specific uses are carried over to the written medium, whereas in HKE the British model is more influential with regard to written data. Because of the different varieties included in the two datasets, these results remain tentative; further confirmation in a more controlled dataset of spoken and written data would be needed.

Secondly, the sample analysis of the verbs *study* and *know* points towards some of the underlying reasons for regional variation. In the newspaper dataset we showed that the use of stative verbs in the progressive is not as influential for variation as has often been assumed; among the top 100 verbs that display regional preferences for the progressive, we found only one member of the group of stative verbs (*appear,* rank 78). In the spoken data, however, this use was more influential, with the verb *know* ranking among the top ten influential verbs. Furthermore, we identified regional variation in the use of the verb *study*; here a different collocational profile between the varieties showed that speakers of some varieties extend the procedural meaning of *study* to the stative meaning of *being a student*.

5. Discussion and concluding remarks

The results presented in Section 4 allow for a short discussion of our methodology and several concluding remarks concerning regional variation in the use of progressive aspect marking. Firstly, the bottom-up approach of creating progressive profiles in order to generate groups of varieties that display homogeneous preferences for progressive aspect marking seems to generate valid results. Especially with regard to the newspaper corpora, the results of the cluster analysis were in line with earlier studies of morphosyntactic linguistic variation: while most South Asian varieties of English are relatively homogeneous, Sri Lankan English seems to be closer to its historical input variety British English (cf. Schilk et al. 2012). With regard to the Southeast Asian varieties our cluster analysis also confirms the idea that Hong Kong English and Singapore English are at different stages of their developmental process, with SinE being the endonormatively stabilized variety and HKE being more strongly influenced by BrE. In this case, somewhat unsurprisingly, regional proximity is far less influential than developmental factors such as linguistic identity construction of the speakers.

In the second step of our analysis, we identified those verbs that are most influential for the differences in the use of the progressive. Owing to the pilot nature of this study it was not possible to perform a comprehensive qualitative analysis of all verbs that were identified by this procedure. However, the analysis of a number of influential sample verbs has shown that the differences in use are not merely quantitative in nature but that there are also new form-meaning mappings to be identified in the use of the progressive in some of the varieties under scrutiny. A case in point is the use of the double or stacked progressive, with two progressive forms following each other. In this case we assume that it is possible that speakers of some varieties (in this case Indian English) may use progressive forms as quasi-nominal substitutes for noun phrases. Due to the limited amount of examples, we can, however, not make any strong claims on the typicality of this phenomenon. It may well be that these uses are highly peripheral and marked even within the respective variety. A further example of usage extension of the progressive can be shown in the wider collocational range of the verb *study* that speakers of HKE and IndE use. The extension to collocates denoting places adds a stative meaning to the progressive use of *study* that is not found in our ENL data.

Finally, our analysis has shown that the use of so-called stative verbs with progressive aspect marking is relatively rare in the written data but more frequent in spoken language. This result shows that the reasons for differences in the use of the progressive may be far more diversified than has sometimes been assumed in earlier research. In spoken language a lectal cline of the speakers may be more influential for differences in language use, whereas in written language the reasons for different preferences in the use of the progressive may go well beyond simple notions of second-language speaker 'mistakes' or linguistic interference.

Notes

[1] The functions of progressive we list are by no means exhaustive. They are examples that simply serve as the illustration of prototypical functions the progressive construction may carry out.

[2] For a comprehensive description of the CLAWS tagger and tagset, see Garside 1987.

[3] Only -ing participles of lexical verbs (tagged _VVG in the CLAWS tagset) were used in the regular expression. Forms of catenative -ing participles (e.g. BE going to, tagged _VVGK) were excluded.

[4] Generally speaking, "[t]he clustering criterion is based on the error sum of squares, E, which is defined as the sum of the squared distances of individuals from the centre of gravity of the cluster to which they have been assigned. Initially, E is 0, since every individual is in a cluster of its own. At each stage the link created is the one that makes the least increase to E" (Upton and Cook 2008).

[5] "Whereas in the 2nd article, it says that the economy is fast rising ever since the Ramos Administration started" (ICE-Phil w1a-011) (Hundt and Vogel 2011: 158).

References

Bhat, D. N. S. (1999), *The Prominence of Tense, Aspect and Mood*. Amsterdam: Benjamins.

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

Brown, R. (1973), *A First Language*. Cambridge, MA: Harvard University Press.

Collins, P. (2008), 'The progressive aspect in World Englishes: a corpus-based study', *Australian Journal of Linguistics*, 28: 225-251.

Collins, P. and X. Yao. (2012), 'Aspects of the verbal System of Malaysian English and other Englishes', *3L: The Southeast Asian Journal of English Language Studies*, 19: 93-104.

Cook, V. (1993), *Linguistics and Second Language Acquisition*. New York: St. Martin's Press.

Das, D. (2010), 'The uses and distribution of non-progressive verbs in progressive forms: a corpus-based study'. Paper presented at the *26th Northwest Linguistics Conference,* Vancouver, May 8-9.

Elsness, J. (1994), 'On the progression of the progressive in Early Modern English'. *ICAME Journal*, 18: 5-25.

Filppula, M., J. Klemola & H. Paulasto. (2009), 'Digging for roots: universals and contacts in regional varieties of English', in: M. Filppula, J. Klemola and H. Paulasto (eds) *Vernacular Universals and Language Contacts –*

Evidence from Varieties of English and beyond. New York: Routledge. 231-259.

Garside, R. (1987), 'The CLAWS word-tagging system', in: R. Garside, G. Leech and G. Sampson (eds) *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Gass, S. M. and L. Selinker. (2001), *Second Language Acquisition: An Introductory Course*. London: Lawrence Erlbaum.

Haspelmath, M. (1999), 'Why is grammaticalization irreversible?'. *Linguistics,* 37: 1043-1068.

Housen, A. (2002), 'A corpus-based study of the L2-acquisition of the English verb system', in: S. Granger, J. Hung and S. Petch Tyson (eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins. 77-116.

Hundt, M. and K. Vogel. (2011), 'Overuse of the progressive in ESL and learner Englishes – fact or fiction?', in: J. Mukherjee and M. Hundt (eds) *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap.* Amsterdam: Benjamins. 145-166.

Jenkins, J. (2000), *The Phonology of English as an International Language.* Oxford: OUP.

Kachru, B. (1983), *The Indianizazion of English*. Dehli: OUP.

Kachru, B. (1992) *The Other Tongue: English across Cultures, 2nd edition.* Urbana, IL: University of Illinois Press.

Mair, C. (1998), 'The corpora and the major varieties of English: issues and results', in: H. Lindquist, S. Klintborg, M. Levin and M. Estling (eds) *The Major Varieties of English: Papers from MAVEN 97*, Växjö 20-22 November 1997. Växjö: Växjö University Press. 139-157.

Mair, C. and M. Hundt. (1995), 'Why is the progressive becoming more frequent in English?', *Zeitschrift für Anglistik und Amerikanistik* 43: 111-122.

Mair, C. and G. Leech. (2006), 'Current change in English syntax', in: *The Handbook of English Linguistics*. Blackwell: Oxford. 318-342.

Mesthrie, R. (2004), 'Synopsis: Morphological and syntactic variation in Africa and South and Southeast Asia.', in: B. Kortmann and E. Schneider in collab. with K. Burridge, R. Mesthrie and C. Upton (eds) *A Handbook of Varieties of English. Volume 2: Morphology and Syntax*. Berlin: Mouton de Gruyter. 1132-1141.

Mesthrie, R. and R. M. Bhatt. (2008), *World Englishes: The Study of New Linguistic Varieties*. Cambridge: CUP.

Mukherjee, J. and S. Hoffmann. (2006), 'Describing verb-complementational profiles of new Englishes: a pilot study of Indian English', *English World-Wide*, 27: 147-173.

Nesselhauf, N. (2007), 'The spread of the progressive and its 'future' use', *English Language and Linguistics,* 11: 193-209.

Platt, J., H. Weber and M. L. Ho. (1984), *The New Englishes*. London: Routledge.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Ranta, E. (2006), ‘The ‘attractive’ progressive – Why use the *-ing* form in English as a lingua franca?’, *Nordic Journal of English Studies,* 5: 95-116.

Rogers, C. K. (2002), ‘Syntactic features of Indian English: an examination of written Indian English’, in: R. Reppen, S. Fitzmaurice and D. Biber (eds) *Using Corpora to Explore Linguistic Variation.* Amsterdam: Benjamins. 187-202.

Römer, U. (2005), *Progressives, Patterns, Pedagogy. A Corpus-driven Approach to English Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: Benjamins.

Schilk, M. (forthcoming), ‘Using currency annotated lexical and POS-tag profiles for the study of linguistic variation: a data-exploration of the International corpus of English.’

Schilk, M., T. Bernaisch and J. Mukherjee. (2012), ‘Mapping unity and diversity in South Asian English lexicogrammar: verb-complementational preferences across varieties’, in: M. Hundt and U. Gut (eds) *Mapping Unity and University Worldwide: Corpus-Based Studies of New Englishes.* Amsterdam: Benjamins. 137-166.

Schmied, J. (1994), ‘Syntactic style variation in Indian English’, in: G. Blaicher and B. Glaser (eds) *Anglistentag 1993 Eichstätt proceedings*. Tübingen: Niemeyer. 217-232.

Schneider, E. W. (2007), *Postcolonial English: Varieties around the World.* Cambridge: CUP.

Sharma, D. (2009), ‘The typological diversity in New Englishes’, *English World Wide*, 30: 170-195.

Smith, N. (2002), ‘Ever moving on? The progressive in recent British English’, in: P. Peters, P. Collins and A. Smith (eds) *New Frontiers of Corpus Research: Papers from the Twentieth First International Conference on English Language Research on Computerized Corpora*, *Sydney 2000*. Amsterdam: Rodopi. 317-330.

Smith, N. (2005), *A Corpus-based Investigation of Recent Change in the Use of the Progressive in British English*. Unpublished PhD thesis, University of Lancaster.

Smitterberg, E. (2005), *The Progressive in 19th-century English. a Process of Integration*. Amsterdam: Rodopi.

Suzuki, R. and H. Shimidora. (2006), ‘Pvclust: an R package for assessing the uncertainty in hierarchical clustering’, *Bioinformatics*, 22: 1540-1542.

Swan, M. and B. Smith. (2001), *Learner* English *– a Teacher's Guide to Interference and Other Problems*. Cambridge: CUP.

Upton, G and I. Cook. (2008), *A Dictionary of Statistics (2nd rev. ed.)*, Online version. Oxford: OUP. EISBN: 9780191726866.

Van Rooy, B. (2006), ‘The extension of the progressive aspect in Black South African English’, *World Englishes*, 25: 37-64.

Ward, J. H. (1963), ‘Hierachical grouping to optimize an objective function’, *Journal of the American Statistical Association* 58: 236-244.

Wee, L. (2004), 'Singapore English: morphology and syntax.', B. Kortmann and E. Schneider in collab. with K. Burridge, R. Mesthrie and C. Upton (eds) *A Handbook of Varieties of English. Volume 2: Morphology and Syntax*. Berlin: Mouton de Gruyter. 1058-1072.

Williams, J. (1987), 'Non-native varieties of English: a special case of language acquisition', *English World Wide*, 8: 161-199.

Neology: from word to register

Antoinette Renouf

Birmingham City University

Abstract

In this paper, we investigate the context within which a neologism occurs. A pilot study of a diachronic journalistic corpus confirms that a coinage or new word formation which names or is associated with a major topical event will often not occur in isolation, but become part of a communal and cumulative activity. Further novel language use will emerge and 'converge' at around the same time, as will lexical, semantic and grammatical variants of existing words. If the new real-world area of concern is sustained in the media, these words and phrases begin to co-occur, forming a loose inter-collocational network which we deem to be an incipient 'register'. The paper will provide data centred on neologistic activity in UK journalism in mid-late 2011, reflecting the response of the UK leadership to the national economic crisis, and its ripple effect through social institutions and the media. The findings may serve to alert English language practitioners to the merits of examining the larger context of a neologism to discover further interrelated new words, and novel ways of representing lexical information.

1. Introduction

This paper forms part of a larger study which takes a corpus-based approach to the diachronic study of a very large corpus of mainstream UK news texts dating from 1984-2012, to arrive at a progressively finer definition of English neology. Previously, we have investigated the life-cycle of a neologism beyond its first occurrence in text (Renouf 2012a), as well as the nature of hapax neologisms (Renouf forthcoming). This approach comes into its own at its intersection with the language professions: specifically lexicography, terminology and translation studies, where it provides knowledge crucial for language-descriptive and strategic purposes (Renouf 2012b).

In this paper, we depart from the language-descriptive tradition of representing neologising as an isolated event. In the real world, events have consequences, and spawn and interact with other events, whether in linear, radial, hierarchical or other patterns. In journalistic text then, which mirrors closely the real world, a similar pattern of word creation and use can be expected.

Our data observation has led us to the belief that a host of neologisms and novel language use can emerge within a similar period of time. In this paper, we aim to confirm the hypothesis that neologising can be a communal and cumulative activity, and to propose a provisional set of categories of novel behaviour. We cannot guarantee a representative or exhaustive categorisation, given the scale of corpus data. Our sample words are those which, in addition to contributing to aspects of the topic at hand, show significant growth in frequency within the chosen period of study, 2009-2012, centring on mid-late 2011.

We investigate the way in which a key, or 'alpha' neologism (see Definitions in 1.1 below), is bolstered, during its period of topicality, by other words and phrases. These formations are a combination of formal, semantic and grammatical neologisms, new or existing synonyms, inflectional and derivational innovations, morpho-syntactic variants of existing words; and other elements required to present or comment on new information, or known information in a novel way. Some formations are already doing a job somewhere in the larger corpus, and are temporarily diverted and modified to contextualise the new words. These may simply be minor players in text which are temporarily revived and increased in frequency; or they may be borrowed from a specialised domain of technical/foreign language.

We shall provide illustrative data from a period of dynamic neologistic activity in UK journalism, which took place within the chosen time-frame, concerning the UK's national economic crisis, the responses of its political leadership, the responses of the opposition parties, and the consequences for domestic policy and daily life. Our source data are diachronic, but our focus is just on this recent period in UK history.

1.1 Definitions of key terms and concepts

"Neologism"

Finding suitable terms for novel words and uses is problematic in this paper. For one thing, we were looking at words which may be new, but which could also just be 'converging' or shifting into the new domain from one in which they were fully assimilated; or could be reviving after a period of disuse. For another, we are not concerned with the ultimate fate, assimilation or otherwise, of the neologisms and new uses we find in the chosen period of journalistic activity, but only in how they emerge and interact within this interval in time.

The general view is that "a neologism is a word which becomes part of the language" (Bauer 2001: 39), though Bauer himself is of the view that "it is probably not possible to tell at the point when a word is coined whether it will turn out to be a 'nonce' word or a neologism in this sense" (2001: 39-40). Bauer continues that "a term is required which is neutral with regard to…diachronic implications…", and he proposes the alternative term 'coinage', found in Marchand (1969: 9) and Strang (1970: 27).

We ourselves use the terms 'neologism' and 'neologistic' to refer generally to the novel words and uses which appear or converge on our data during the chosen time-span of text. We may use the terms 'coinage' or 'coining' to emphasise that a form has, as far as we can tell, been newly invented.

"Register"

In this paper, we illustrate how a new word can become bound up in a swirl of lexical innovation and change in text, and we tentatively use the term 'register' –

possibly not more than an 'incipient' register – to characterise that situation. It amounts to a particular set of words and phrases coming together to characterise (and report on) reality at a certain time, attracting to them particular collocates, synonyms, and semantic and grammatical features.

While there is little overall consensus in defining the term 'register', the notion of permanence in a register is pervasive in the literature: Halliday (1964) and Halliday and Hasan (1976) have in mind an established pattern of language, rooted in a 'typical' context of situation. Ferguson (in Biber and Finegan 1994: 20) says that "people participating in recurrent communication situations tend to develop similar vocabularies, similar features of intonation, and characteristic bits of syntax and phonology that they use in these situations"; Crystal (1991: 295) defines register as "a variety of language defined according to its use in social situations"; while Biber and Finegan (1994: 5) state that "register is a language variety viewed with respect to its context of use", though Biber does allow of changing parameters over time. Our definition differs from all these in that we are contemplating:

a) a potential register at its birth – a fragmentary constellation of particular words, inspired by a unique, large-scale event, and
b) an ephemeral phenomenon, where a particular real-world event temporarily sparks neologistic activity.

Halliday and Hasan's definition of 'field' is "the total event, in which the text is functioning, together with the purposive activity of the speaker or writer; it thus includes the subject-matter as one element in it" (1976: 22). This also applies to our notion of register. We can say that the 'field' to which our new words and structures relate is the political situation surrounding the UK general election of 2011, within the social context of a national and global economic crisis. Their purpose is to persuade and motivate the electorate, by dint of euphemism and other rhetorical devices. The particular topic or subject-matter is basically society, and how it can achieve more capacity and efficiency, whilst receiving increasingly lower public funds. The tenor of our data meanwhile comprises politicians or spokespeople for government and public bodies and services, quoted directly or indirectly by journalists. The mode involved is the written word of journalistic text. Since our focus is on the language generated during a period of social crisis, perhaps our register is closer to Wardhaugh's (1986: 22) restricted definition as "a specific set of linguistic terms which can be associated with some external factor".

We also differ importantly from the approaches above in that this study is corpus-linguistic, more specifically diachronic-corpus-based, and not text-linguistic. We study the emergence and convergence of words across the texts of the corpus in the chosen time period.

An important reason for our appropriating the term 'register' is simply that it has a psychological flavour, which chimes with our sense that one joins a sort of mental club when conforming to a register; a particular state of mind is required to apply it consistently.

“Convergence” and “diversion”
To characterise the accumulation of new words and uses observed, we may use the terms *convergence* and *diversion* (and variants). “Convergence” already has several meanings in linguistics (e.g. Giles and Coupland 1991: passim).[1] We use it to mean that various existing words “converge” on the domain of new interest, to join neologisms to elaborate the topic at hand. They are “diverted” from their conventional use and meaning and temporarily inserted into the new domain.

“Alpha” and “beta neologisms”
In our analysis of the neologistic behaviour in the reportage and commentary in UK news text, we differentiate between key new ‘alpha’ neologisms, the ‘drivers’, like *big society*, which characterise and usher in a whole new topic area, and new ‘beta’ terms, like *squeezed middle* and *social cleansing*, which are subsumed under that main ‘alpha’ topic umbrella, subordinate to it in scope of reference, but which spawn their own, more specific, neologistic groupings. The distinction is impressionistic, but it is backed up by criteria of importance in the form of frequency, earliest emergence and persistence, with *big society* holding sway on all measures, *squeezed middle* following, and *social cleansing* in more restricted use. For simplicity’s sake, we have only built ‘alpha’ and ‘beta’ levels into our analytical hierarchy, though the beta terms clearly sit at different levels of influence and specificity. The application of the simple notions of ‘alpha’ and ‘beta’ is experimental; they may after further data analysis be found to conceal a more multi-layered network of word association, both conceptually and temporally.

There is no implication that all incoming or ‘converging’ new words and uses collocate with all key alpha and beta terms, or indeed with all other new words and uses, from the outset. We see, rather, strands of collocation and inter-collocation within the individual topic areas involved, and in time, possibly across them.

1.2 Research background, data and method

Since 1990, we have conducted automated studies of diachronic news text to monitor lexical and semantic change, the earliest being the 1990-1993 AVIATOR project.[2] The 1997-2000 APRIL project[3] monitors morphological productivity by automatically recording and analysing each new word at its first point of entry into our evolving journalistic corpus, and provides the neologistic output of mid-late 2011 (within a 2009-2012 time-frame) which is the focus of this paper.

Meanwhile, the contextualised output for these new words is extracted from the downloaded news corpus and presented in interpretable form by the WebCorp Linguist’s Search Engine (2004-2006) (Kehoe and Gee 2009). The raw data used are the corresponding section of a 1.3 billion-word chronological corpus of UK *Guardian* and *Independent* news texts, published between 1984 and 2012.

We work with the linguistic parameters of first-order collocation, and patterns of frequency; and statistical measures of relative and significant frequency. We have still to investigate the benefits of using a kind of second-order lexical collocation measure (see e.g. Pacey et al. 1998 or Mollet et al. 2011), for logging emerging collocation and inter-collocation within the overall evolving network of convergence.

Contextualised information is presented in the form of concordance lines and collocation analyses. The graphical representations of statistical information are accessed in the form of time-graphs and 'heatmaps'.[4] A time-graph presents the changing frequency profile of a word across time, scaled for time and frequency per million words. It can provide an actual frequency profile, a more interpretable smoothed frequency line if the data are sufficient and stable enough to support it; and statistical information on significant deviation, or change, in the form of a trend analysis. A heatmap is a graphical representation of data where numerical values are represented as colours. It converts the significance scores for each monthly time chunk of lexical data into block colour, and thus shows the changing profile of significant collocates for a given word across time.

The structure for this paper is to select a particular new alpha term, and to identify and classify the neologistic and other targeted language behaviour which accompanies it in creating new journalistic text during a particular period of time.

2. Case studies

This section will present a series of mini-case studies, each based on an alpha or beta neologism and its associated neologisms and convergent activity in text. Most neologisms observed in this period have so far been nouns, as expected in a politically-driven context, though words of other grammatical persuasion certainly do contribute at some level to the debate (see later sections of this paper).

The alpha term we select is *big society*, one which became prominent in 2010, peaking in 2011 (and sustained through 2012). *Big society* is itself a consequence of the social environment and language which preceded it. The national banking crisis of 2008 is its backdrop. Within that earlier context, the alpha term might have been *banking crisis*, and a key beta term *austerity*. In this study, however, we focus our attention on the words and phrases of everyday social upheaval, around 2011.

In this period, the UK political parties went through a period of electioneering and an election. The Coalition government, once elected, was faced with an ongoing debt crisis, and they launched wide-ranging measures to resolve it. *Big society* was the term they selected to characterise their agenda for change to the social fabric of the country. Their overall policy was to increase public performance, and public responsibility for maintaining the country's infrastructure, while reducing government funding to public services. *Big society* thus names an ideological construct.

Whilst *big society* in principle subsumes all areas and vocabularies of public and private domestic life in the UK, our aim is to focus this study on government and public services, and to avoid specialist areas like finance, which would take us down a slightly different track of neologistic activity. The beta term *finance* would pull in a rather specialised network of economic terms: new ones, like *VILE* (‘volatile inflation, little expansion’) (Kitson et al. 2011), and newly-topical ones, like *quantitative easing*, *QE1* and *QE2*, *fiscal*, *inflationary*, and *recession*.

The convergence of neologisms around *finance* and *austerity* overlap with those of *big society* at the point where social consequences result from austerity measures, with neologisms relating to the shared concerns of unemployment, social exclusion, hardening attitudes to immigration, and so on. One imagines a series of overlapping and interweaving universes of neology making up the major themes of UK society in this period.

2.1 Alpha neologism

Big society

Preamble: The term *society* is not new, but since the Conservative Prime Minister (henceforth ‘PM’) Margaret Thatcher’s notorious decree in 1987 that ‘There is no such thing as society’,[5] the term has been a hot potato tossed between successive UK governments. It has remained at a significant level of topicality in the evolving journalistic lexicon. Meanwhile, since 1984, the start of our corpus, *big* has developed a sense other than ‘large in size’; which is a derogatory term generally used by Conservatives (henceforth colloquially referred to as ‘Tories’) to describe a government or public sector considered to be too large and inefficient. This sense was used by Reagan in the US back in July 1986:[6]

(1) 00/07/86 We stand for the opportunities and welfare of ordinary people against **big government**, big business and big unions.

The main political parties were thrashing around for slogans in the run-up to the 2010 national elections, with David Cameron promoting “*people power* not *state power*”, and all parties playing with the idea of ‘big’; the Labour party floating terms including *big government* and *big state*. By 2010, these two had been appropriated by Tories, to criticise Labour policy, after which their frequency profiles rose sharply.

Big society was conceived by the Tory opposition as the inverse of the *small government* policy which was heavily promoted by PM Thatcher in the late 1980s. It enters our data on 19/03/09, with a peak at the start of 2011, and then a drop and a new growth towards the end of 2012. The neologism *big society* is used by the Tories to mean ‘reducing governmental control and giving power to the people’.[7] It is a central plank in the ‘incipient register’ which we have observed in journalism during the period 2009-2012.

The current Tory party first decided to stake all on the notion of *big society* in 2009, in their pre-election speeches. In Figure 1, we see *big society* in early use in 2009, where it is typically written in lower-case, and flagged as novel by the conventions of inverted comma and explanatory gloss (Renouf and Bauer 2001), within contexts showing its meaning to be in clear contrast with Labour policies (and slogans).

19/03/09	He said that as “a progressive Conservative”, he believed in “a **big society**, not a **big state**”
10/11/09	the Conservatives want to create a “**big society**” as an alternative to **big government**.
10/11/09	Conservatives’ plans for a more equal Britain; their “**big society**” to Labour’s “**big state**”
11/11/09	Tories want to create ‘**big society**’, says David Cameron. The era of **big government** is over
16/11/09	Second, to counterpose the “**big society**” to the “**big state**”

Figure 1. Extract of total concordance lines for *big society* in 2009 (5,832 occs.), case insensitive (henceforth CI)

Big society was launched again by the Tories in 2010 in their spring Election Manifesto, when their candidate alternative terms, *social responsibility* and *people power*, were dropped. In the UK General Election on Thursday May 6, 2010, no party gained an overall majority, and *big society* next reared its head in the programme of the resultant coalition government of Tories and Liberal Democrats from May 11, 2010.[8] Thereafter, it became a hot topic, as the surge in frequency of occurrence shows in the time graph in Figure 2.

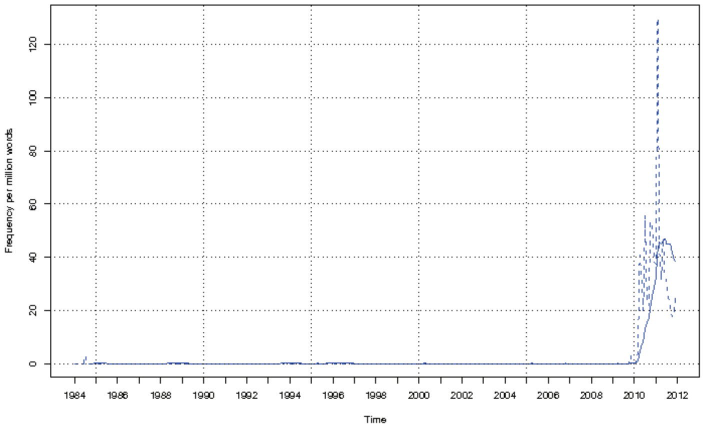


Figure 2. Time graph for *big society*, CI

As 2011 arrived, the Labour leader, Ed Miliband, tried to cast doubt on the Tories’ vision for the *big society* and re-appropriate it for his party, as seen in Figure 3.

15/01/11	The Labour leader is putting together his own alternative to David Cameron's "**big society**".
15/01/11	Miliband told the Labour conference that Labour had to reclaim the "**big society**", an idea in tune with its values, not the Conservatives.

Figure 3. Early 2011 references to the term *big society*

However, the term had been firmly appropriated by the Tories. It began to appear in upper-case, as *big society*, from early 2010. From an early stage, the journalistic use of the term, both in lower and upper case, typically reveals scepticism on the part of both the ordinary public and Tory opponents as to its meaning and value; as shown in Figure 4.

08/02/10	Conrad Black dismisses David Cameron as "an Obama emulator" who "cites only leftists as his intellectual inspiration for what he unpromisingly calls 'the **Big Society**'".
07/11/10	Labour believe the **big society** is a fig leaf for an ideological mission to shrink the state
03/06/11	Whatever that phrase '**Big Society**' might mean, libraries would contribute to its value.
11/09/11	The **Big Society**, adds Joe, is "Thatcherism with a smile".
12/09/11	The concept of the **Big Society** is so empty that universities have put it at the top of their research agenda
12/09/11	RNLI chief executive says **Big Society** gives no added value to charities.
15/09/11	It's emblematic that Mr **Big Society** has become Mr Big Scissors.
18/10/11	It is more proof that the **Big Society** is just Tory-speak for "you're on your own.'"

Figure 4. Extract of contexts for *big society*, Feb 2010 – Oct 2011

And by the end of 2011, the negative consequences of the 'big society' policies were beginning to be clearly articulated. By now, *big society* was frequently missing its definite article, as seen in Figure 5.

04/12/11	**Big Society** millionaire men telling (mostly) poor (mostly) women that they should be doing more for no pay.
12/12/11	an attempt to plug the holes of the CQC[9] by resorting to '**Big Society**'
14/12/11	**Big Society** means Government turns everything into a business.
17/12/11	It is **Big Society** in action. "We believe 750,000 people volunteered and gave food to foodbanks this year."

Figure 5. Dec 2011 extract of contexts for *big society* minus definite article

The overarching goal of the incoming Coalition government was to reduce the national deficit following the banking collapses in 2008-2009. The *big society* policies were accordingly two-fold. Citizens were to be responsible for the welfare of society; while the Coalition implemented public spending cuts and austerity measures. Unsurprisingly, there were consequences, and these were reflected in some of the more successful neologisms in 2011.

2.2 Beta neologisms

As said above, it is useful to see the metaphor *big society* as an ‘alpha’ topic noun in its scope of reference, and the neologisms functioning within its topic orbit as ‘beta’, governing the content of text at the next level of topic specificity. Two key ‘beta’ neologisms which emerged to characterise logical consequences of evolving *big society* policies were *squeezed middle* and *social cleansing*.

Squeezed middle

The coinage *squeezed middle* is based on a metaphor borrowed by the Labour Party from the Clinton administration to undermine the credibility of the Coalition’s *big society* policy, by highlighting the repercussions of public spending cuts and increased taxation for hard-working middle-class citizens. The time-graph in Figure 6 shows this noun phrase emerging haltingly in the pre-election campaign in September 2009 and then soaring in frequency from early 2011.

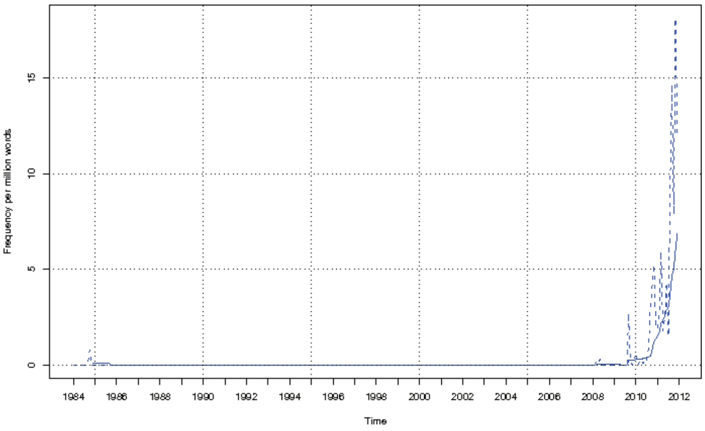


Figure 6. Time-graph for *squeezed middle* (866 occs.), CI

In 2009-2011, however, it seems from our data that the meaning of the term remained generally unclear, as exemplified in Figure 7. Even in late 2010, the current Labour leader, Ed Miliband, himself displayed ignorance of the average salary of the middle class affected, and the term was still being defined by Labour a year on.

27/09/09	Labour's new focus on the "**squeezed middle**", though what that means remains vague.
29/09/09	the "**squeezed middle**" (whatever that means)
28/11/10	Labour today moved to define its principal target group, the "**squeezed middle**", as people on an income of between £16,000 and £40,000-£50,000.
28/11/10	the Labour leader found himself struggling in interviews to explain who were the "**squeezed middle**".
10/01/11	Mr Miliband's team are casting around for a different phrase to replace the "**squeezed middle**" after Mr Miliband appeared to get into a bit of a tangle on the radio over who actually constitutes this group.

Figure 7. Extract of contexts for *squeezed middle*, CI, Sept 2009 – Jan 2011

By the end of 2011, however, it was finally settling into use, as seen in Figure 8 contexts.

03/12/11	only a few months ago phrases like the "**squeezed middle**" were being mocked and now they're very much part of the national conversation.
03/12/11	The longer-term reality: many working families were experiencing tough times before anyone talked about the "**squeezed middle**"
08/12/11	ordinary citizens – especially among the **squeezed middle** classes – are the ones enduring the pain
22/12/11	a promise that people in the **squeezed middle**, defined as a combined income of below £60,000, receive high-quality affordable housing.

Figure 8. Extract of contexts for *squeezed middle*, CI, Dec 2011

Social cleansing

The term *social cleansing* echoes the sense of *cleansing* in the 'ethnic cleansing' of the Kosovo war, used to mean 'the removal of undesirable social elements'. In Kosovo, this sense of *cleansing* evolved to mean 'removal of undesirable ethnic elements', and ultimately, 'genocide'. This neologism has been coined as another response to the *big society* agenda, specifically to the ramifications of the Coalition's cuts to local council budgets. From mid 2009, these austerity measures led richer London boroughs[10] to raze poor accommodation to eliminate poorer residents. This term was used primarily by Labour critics of the measures, as seen in Figure 9.

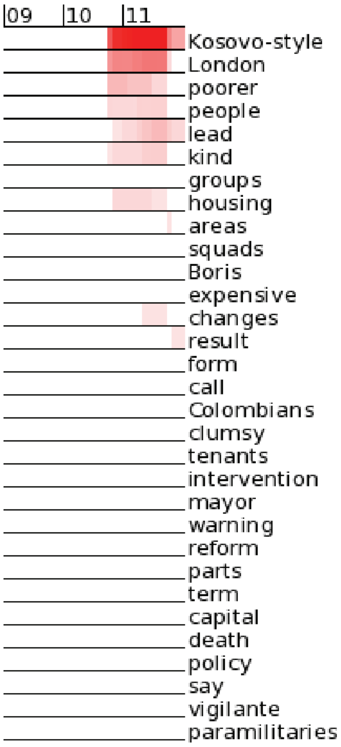
15/07/09	Labour opponents have alleged that 'Decent Neighbourhoods' is designed to release the Council from its obligation to house people on low incomes and to see large numbers of those Labour-voting people, move out of the borough altogether. Andrew Slaughter MP has called it a policy of "**social cleansing**".

24/02/10	The Labour group thinks changes give a green light to the more radical proposals of Hammersmith and Fulham to create what they call “decent neighbourhoods” in place of existing social housing estates and what their critics say would be the covert “**social cleansing**” of poorer residents and their Labour-voting tendencies.

Figure 9. Defining contexts for *social cleansing* in 2009 and 2010

The debacle brewed in the wings of topicality until 2011, when some London boroughs adopted the practice of meeting their housing obligation to poorer citizens by dispatching them to rented accommodation in cheaper parts of the country, as outlined in the contexts in Figure 9. On Nov 2, 2011, Boris Johnson, the Tory Mayor of London, finally hit the news when he brought to the term *social cleansing* a connotation of *ethnic cleansing*, by saying: “We will not accept a kind of Kosovo-style social cleansing of London…” This was a calculated embarrassment to the Coalition government, from whom he sought to distance himself. The heat map for *social cleansing* in Table 1 shows the late 2010 collocational shift towards the London scandal.[11]

Table 1. Heatmap for *social cleansing* (138 occs.), CI



2.3 “Reactive neologising”

Reactive neologising is a striking feature of the journalism leading up to the 2010 general election until end 2012, and particularly in 2010-2011. The period was extremely politically driven, with daily reportage on the jockeying for position by the main protagonists. The manoeuvring led to a high degree of political posturing via neologisms of a particular kind: emblematic nouns, characterising shifting ideologies and positions between the parties and politicians. One side coins a term to characterise a politically strategic concept or phenomenon; the opposition coins a second term which attributes to the concept a negative connotation. Though they are loosely co-referential, it is a stretch to call these attitudinally opposing pairs synonyms.

To begin by scrutinising the three key alpha and beta terms introduced above, we see that each goads the other political camp into coining one or more “semantic counterparts”. Just as in 2009, shown in Figure 1, Labour’s *big state* and *big government* provoke the Tories into countering with *big society*; in 2010/11, *big society* begets *small society*; *squeezed middle* engenders *alarm clock Britain*; *regeneration* leads to *social cleansing*; *localism* spawns *nimbyism*, *nimbyism* licences *imbyism*, and so on. These new pairs are also found to collocate with each other, strengthening the evolving lexical framework, as illustrated in the concordanced output in Figures 10-14.

Of course, these are highly-motivated but not exclusive pairings: *regeneration*, for example, also collocates significantly with *big society* (28 times within a 6-word span; 40 in a 10-word span; 57 in a 15-word span; 858 times within a document span).

Small society* in response to *big society

In the case of the Tory *big society*, we have previously cited the opposing term *big government* piloted by Labour. Another rival is *small society* (68 occs., of which 4 upper-case), a Labour term used to re-cast *big society* as really being a small society of local neighbourhood groups carrying civic and social responsibility without reciprocal state funding. The close proximity of these opposing terms shown in Figure 10 makes the meaning of *small society* clear.

13/05/10	Whatever happened to the **Big Society**? At this week’s coalition marriage, the **small society** was much in evidence.
12/06/10	not the **Big Society**, but the **Small Society**, as Dave might put it.
19/07/10	“I take the **big society** seriously,” Miliband said. “But it is a piece of doublethink; a **small society** maintained by voluntarism and charity. I want a **bigger society**, based on reciprocity...”
27/08/10	the influence of the individual citizen, allegedly a key element of the “**big society**” – in reality the **small society** – is difficult to discern in the policing and health structures
20/02/11	When Cameron refers to the **big society**, he really means the **small society**

29/09/11	his theory of the **smaller society** taking care of the **big society**
26/03/12	it could be about different things: the **big society**, the **small society**.
28/08/12	My **small society** will get by; **big society** won't and shouldn't

Figure 10. Extract of contexts for *small society* in collocation with *big society*

Alarm clock Britain in response to *squeezed middle*

The neologism *alarm clock Britain* (total 77 occs., of which 63% unhyphenated and lower-case) was coined by the Liberal Democrat leader, Nick Clegg in early 2011 to reclaim the territory of the earlier Labour term, the *squeezed middle*. By "alarm clock Britain", he meant "basic rate taxpayers", identified as the key group in the political battle over the government's austerity measures. His aim was to portray the Coalition as being sympathetic to their dilemma. A search on the string *alarm#clock Britain* (where '#' stands for a space, hyphen or nothing) yields the examples in Figure 11 which show that they were essentially characterising the same section of society.

11/01/11	Nick Clegg will speak up for "**alarm clock Britain**". He has been working on his version of Ed Miliband's "**squeezed middle**" for months
16/01/11	Clegg appealing to "**alarm clock Britain**", Miliband to "the **squeezed middle**"
19/03/11	Presumably in a separate universe to "the **squeezed middle**" or "**alarm clock Britain**".
21/03/11	"**alarm clock Britain**", Nick Clegg's description of the **squeezed middle**.
24/03/11	I will help the **squeezed middle** of **alarm clock Britain** by reducing fuel duty
23/11/11	To be widely confused with: **alarm clock Britain**. Do.say: "Le **squeezed middle**, c'est moi."
06/12/11	described as "**alarm clock Britain**" by Nick Clegg, and the "**squeezed middle**" by Ed Miliband.
26/01/12	Whether you call them the '**squeezed middle**', 'hard-working families', or '**alarm clock Britain**', it's the people whose incomes are too high for welfare benefits, but too low to provide financial security.

Figure 11. Contexts for *alarm#clock Britain* with *squeezed middle* (9-word span)

Social cleansing in response to *regeneration*

Social cleansing was the neologism coined by Labour in 2010 in response to the existing Tory term *regeneration*, for a policy of inner-city demolition of poorer housing, to create 'decent neighbourhoods'. *Regeneration* (15,247 occs., CI) is an existing term which fell in frequency from 1990s peaks, but climbed back up in 2011-2012. Implicit in the Tory demolition programme was the removal of poorer people from areas with private housing development and 'gentrification' potential, and the new sense of *social cleansing* refers specifically to the removal of these 'problem families'. In coining *social cleansing*, Labour aimed to expose *regeneration* as a cheap euphemism, and their efforts are reflected in the examples in Figure 12.

22/11/06	What is trumpeted as "**regeneration**" in the local press, has resulted in the **social cleansing** of sections of the city centre,
23/10/10	we encounter "**class cleansing**" disguised as urban **regeneration**
24/06/10	She was active in resisting the **regeneration** of the area, incensed that local families were being dispersed. It was "**social cleansing**".
11/01/11	Fulham's **regeneration** polices have been at the heart of bitter political struggles, with the Labour MP Andy Slaughter condemning them as a form of "**social cleansing**"
18/06/12	One block of maisonettes bore the graffitoed words "**Regeneration** is **Social Cleansing**".

Figure 12. Contexts for *social/class cleansing* collocating with *regeneration*

Nimbyism in response to *localism*

Localism (3,474 occs.; 70% lower-case) was a term revived by the Tory-led coalition, meaning 'community involvement in local affairs, as a substitute for state funding'. It rose in frequency from 2010 and peaked in 2012. *Nimbyism* (374 occs.; 60% lower-case), with as its root the acronym of 'Not in My Back Yard', was coined in the late 1980s, to characterise the self-interested opposition by residents to inconvenient developments proposed for their neighbourhood.[12] It was a term Labour opponents revived to point to the protectionism *localism* was prone to inspiring among Tories, and it experienced a corresponding rise in 2012. Figure 13 presents them together.

08/06/11	one politician's **localism** can be another's **nimbyism**.
10/06/11	just as **localism** is designed to strengthen community involvement, so too it risks becoming a charter for **nimbyism**.
11/07/11	The **localism** bill should be amended to ensure housebuilding is not restricted by **nimbyism**
13/07/11	If **localism** and the big society are to succeed it has to be on the basis of a can-do mentality, rather than **nimbyism**.

Figure 13. Extract of contexts for *nimbyism* in collocation with *localism*

Imbyism in response to *nimbyism*

Imbyism[13] (3 lower-case occs. only) emerges in 2011 as another counter-concept to *nimbyism*. It is used by Labour sympathisers in response to the negative connotation of *nimbyism*, the protectionist attitude amongst Tories, both politicians and private citizens with vested interests. (The author of the latter two instances shown in Figure 14 is the same.)

09/06/11	I look forward to the localism bill, translating “fierce social protectionism” from **nimbyism** (not in my backyard) to “**imbyism**”, a can-do approach built on the back of new neighbourhood plans.
28/11/11	It is encouraging to see local expression of “**imbyism**”, as the neighbourhood plan commits to 2,000 new homes
15/11/12	A chink in the **nimby** armour of the great middle classes is the fact that their children and grandchildren are unlikely to afford to live in the village that they were born in. I would suggest this is an opportunity to promote “**imbyism**”: yes, do build in my back yard.

Figure 14. Contexts for *imbyism* collocating with *nimbyism*

2.4 Existing terms with increased frequency

As we demonstrated during the AVIATOR project in 1990-1993 (Renouf 1993), the emergence of new words is accompanied by a sudden increase in the frequency of some existing words which have a bearing on the new topic. During mid to late 2011, a series of existing managerial terms each grow in frequency, boosting the prevailing *big society* preoccupation with social responsibility, the improvement of performance, and compliance with regulations. They include *scrutiny*, *challenge*; the derivational pairings *sustainable/sustainability*, *accountable/accountability*, *transparent/transparency*, *target/targets*; and the lemmas *empower** and *compliance**.

Some of these terms come directly from the Coalition election manifesto in 2010. *Responsibility* grows only in capitalised form, in the context of *social responsibility*, an early Tory term, later downplayed. *Transparency* reflects one of the stated priorities in the manifesto, namely ‘open and transparent government’, and *transparent* grows rapidly, in part within the emergent term *transparent government*, from this point on, as shown in Figure 15.

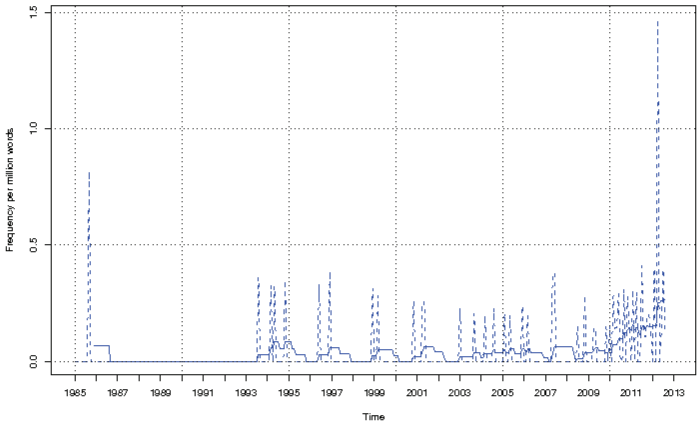


Figure 15. Time graph for *transparent government* (75 occs.), CI

Empowerment, including its derived forms, is another key term in the manifesto, denoting “methods used to increase the power of individuals and communities”. Its collocation with these terms and with *big society* reflects its rhetorical remit in Figure 16.

14/04/10	insight into **empowering** neighbourhoods informed the **Big Society** document
07/07/10	the government’s vision for a **big society**: **empowered** communities and public services delivered by citizen-focused, third-sector organisations.
23/07/10	the “**big society**” was an attempt to **empower** local individuals
22/09/10	this is a **big society** issue – **empowering** the grass roots
23/12/10	the language of popular **empowerment** (“**big society**” being the most obvious example)
22/02/11	the so-called **big society** will not **empower** or strengthen communities
24/02/11	the ‘**big society**’ drive, designed to **empower** communities

Figure 16. Extract of contexts for *regeneration* in its period of growth

The rapid growth of *empower** in the period is seen in Figure 17, where an optional ‘trend line’ in the graph extending beyond the boundaries of the ‘funnel’ of expected variation indicates significant change.

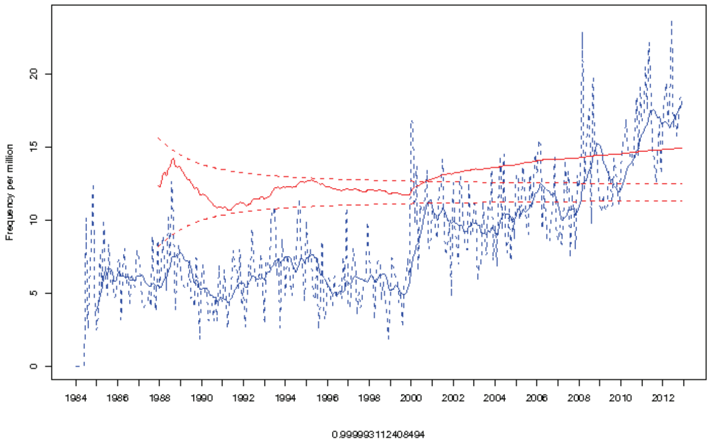


Figure 17. Time graph for *empower** (13,182 occs.), with trend line for significant change, CI

2.5 Existing terms migrating from specialised domains

The ‘alpha’ neologism *big society* needs some bolstering of its topic coverage and rhetorical force. This reinforcement comes not only from ‘beta’ neologisms, but from existing terms migrating into general news text from specialised domains, bringing with them relevant nuances.

On my watch

The phrase *on my watch* (or more precisely [*on* + PossAdj + *watch*]) migrated from military contexts. The ‘watch’ refers to a period of hours of duty during which an appointed serving member of the forces is responsible for the safety of others. It has the rhetorical power carried by the allusion to duty in life and death situations. The earliest occurrences in our data are in fact quotations of the US President, Ronald Reagan, as cited in Figure 18:

00/10/84	‘Presidential leadership means being accountable for events that occur **on your watch**’.
00/03/87	The President said he found the secret bank accounts and diverted funds ‘personally distasteful,’ and acknowledged, ‘as the Navy would say, this happened **on my watch**’.

Figure 18. Early contexts for on [{PP$}|{POS}] *watch* (1,215 occs.), CI

But the phrase gradually became a favourite with British public figures, notably politicians, journalists, bankers, police and the armed forces, to assure people of their reliability and to deflect any suspicion of wrong-doing; but also with the UK media to report on official incompetence and wrong-doing. Like much language use of US origin, it entered the UK news realm via the carrier pigeons of foreign correspondence, and by journalists quoting US usage. It is used by politicians of both parties. The peaks of frequency from 2009 onwards primarily record the fall-out from the Coalition government’s *big society* domestic policies, which are identified in Figure 19 (though *on my watch* and variants also refer to the issues of phone hacking, police corruption and in sport which were occurring in the background at the time; *on my watch* is also a favoured phrase of the Mayor of London, Boris Johnson).

21/09/11	Labour will never tolerate inequality **on its watch**.
26/09/11	**on his watch** NHS trusts are going to face a financial and clinical crisis.
17/10/11	People must judge how this has been allowed to happen **on David Cameron’s watch**.
24/10/11	Boris Johnson: big increase in people injured by knives **on his watch**
02/11/11	**On my watch**, Labour will be a constructive opposition
16/11/11	youth unemployment is one of the biggest crises **on Cameron’s watch**.
23/11/11	a lost generation of young people. It’s happening **on your watch**
28/11/11	**On his watch** patients are waiting longer for treatment

Figure 19. Contexts for on [{PP$}|{POS}] *watch* (1,215 occs.), CI

Strangely enough, the alternative phrase [*under* + PossAdj/N + *watch*] appears regularly in our data from 1995, and contributes a further 370 occurrences, a variant not attested in dictionaries. Of these, 73 are possessive proper nouns.

Outreach

This compound noun (which occurs 99.5% un-hyphenated, 90% in lower-case, 10% as PN of organisations) means "services provided to populations which might not otherwise have access to them". In our earliest data, it refers to foreign aid programmes in former colonies and overseas locations. But since mid 2011, it has been increasingly adopted to refer to a typical *big society* goal – that is, "peripatetically reaching out to segments of the community by use of awards and grants (i.e. not through development of infrastructure)". Its recent remit is illustrated in the extract of contexts in Figure 20, which shows that even these short-term attempts to stretch UK services become subject to cuts.

12/07/11	hospital psychiatric liaison teams, care home outreach teams and palliative care outreach teams, should share expertise
13/07/11	an untested notion that outreach and bursaries will cover a vast structural bias against the poorest from studying.
19/07/11	outreach programmes that send out graduates into their own communities
25/07/11	I got involved in our outreach activities, as schools liaison officer
26/07/11	adult guidance services, fostering community outreach work,
02/08/11	with many areas seeing outreach schemes close altogether.
10/08/11	the government cut Haringey's budget by £45 m, which resulted in huge cuts in crucial outreach youth services

Figure 20. Contexts for search term *out#reach* (3,421 occs.), CI

Stakeholder

There is a phrase in general use, 'to have a stake in something', where 'stake' means 'interest' or 'involvement', but the term *stakeholder* itself was originally a term from corporate business, meaning 'investor'. Its gradual move into the political and social mainstream of language use was fostered by Tony Blair during his premiership. On 07/01/96, Blair first mentioned the word *stake* to characterise the relationship a UK citizen has to public life. On 08/01/96, he first referred to the UK as a *stakeholder economy*. In 2002, he introduced the notion and reality of a *stakeholder pension* (i.e. a riskier one, whereby rates were not guaranteed). By 2011, with a Tory-led regime, the corpus shows that all UK citizens are now *stakeholders*, sharing in the financial risks (and theoretically benefits) of the country. This is political manoeuvring, whereby ideological shifts are imposed on society by new terminology from other domains, typically bureaucratic, corporate and here, financial. This previously occurred in the 1990s, under Tory PM John Major, when hospital patients, airline and train passengers, and recipients of other services, were all re-designated '*customers*'. *Stakeholder* goes a step further, casting the citizen in a monetary light.

06/07/11	discussions with **key stakeholders** and **interested parties** in an attempt to sell the **businesses**
07/07/11	A **key** challenge facing **social entrepreneurs** is the need to measure social impact and answer to **stakeholders**.
07/07/11	amazing feedback from **our patients** and other **stakeholders**
08/07/11	the **marketing of all products and services**, across all our **key audiences** and **stakeholders**.
08/07/11	tell your **stakeholders** what your **charity** has achieved this year
10/07/11	we had to ensure our **members**, **players**, **supporters** and all other **stakeholders** could see we had addressed the issues
11/07/11	**Higher education institutions** will struggle in the marketplace unless they make sense to **stakeholders**.

Figure 21. Contexts for *stake#holder*/*stake#holders* (10,304 occs), CI

Searching on the (case-insensitive) string *stake#holder*/*stake#holders* (of which 99.4% of occurrences are unhyphenated; 95% lower-case), our corpus data indicate that from mid 2011, *stakeholder* means 'a person, organisation or group who affects, or can be affected by, an organization's actions'; or a 'member'. But it also means 'someone whose favour needs to be retained by a public body or company through communication'. In the examples in Figure 21, the agents are clearly social entrepreneurs, public administrators and corporate people. But the term is likely to broaden its remit.[14]

2.6 Semantic neology

Existing words pulled into the orbit of an alpha neologism like *big society* from other textual domains may go further than bringing their specialised nuances, and be given a new, purpose-built sense.

Engage with

An example of contemporary semantic evolution can be seen with the established phrasal verb *engage with* and its inflections. It has evolved into one of a number of managerial terms, in which guise it has increased in frequency of use, as shown in Figure 22.

In our earliest corpus data, *engage with* and its inflections are used in technical and military contexts, as illustrated in Figure 23.

The phrasal verb *engage with* and inflections also show early use in reference to deep, philosophical, moral and practical problems and conditions, as seen in Figure 24. Where a moral context was involved, *engage with* was accompanied by a noun denoting a thorny issue, and the verb implied that somebody was taking on a degree of responsibility, that there was a real commitment to resolving that issue.

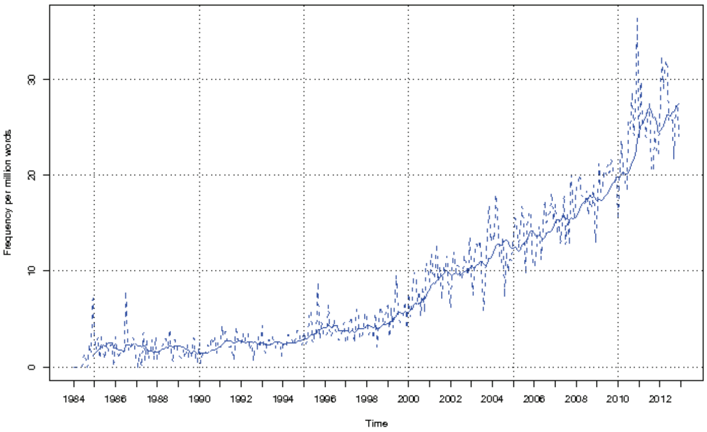


Figure 22. Time graph for the search term *engag* with*

Technical
00/11/85 the enzyme’s surface **engages with the substrate molecule**.
Military
00/06/88 Israeli military are **engaged with the South African military**

Figure 23. Earlier uses of *engag* with* in technical and military contexts

Moral/philosophical
00/11/84 the difficulty in **engaging with** this argument is that religious language has decayed
00/07/86 it first seriously **engaged with** the problems of macro-economic policy’
00/11/86 Professor Lifton’s psychiatry is deeply **engaged with** the causes and effects of barbarity
00/09/88 a determination to **engage with** weighty political and psychoanalytic themes
05/11/88 he is closely **engaged with** the notion of authorial identity
08/04/90 such facets of Christianity must **engage with** global concerns

Figure 24. Earlier uses of *engag* with* in moral and philosophical contexts

By mid 2011, however, we see a peak in the use of *engag* with* in a new, looser, weaker sense. Now it simply means to ‘involve/meet/contact/communicate with/be interested in’, and even ‘get on well with’. Its new collocates are now ‘points of social contact’, such as *communities*, *audience*, *public*, *professionals*, *stakeholders*, *sector*, as seen in Figure 25. But the original semantic weight of the verb is appropriated, to dignify what are now more basic activities, requiring less abstract cogitation.

05/07/11	We need to ensure an **engagement** with **stakeholders** and the **public**
21/07/11	Politicians wanting to find new ways of **engaging** with the **public**
08/08/11	The lack of funding makes it difficult for venues to **engage** with their **communities**
24/08/11	There are many ways for the private sector to **engage** with these **community** needs
31/08/11	How key public services **engaged** with **communities** before, during and after the riots.
10/10/11	relations improved by simply talking and **engaging** with the **sector**

Figure 25. Contexts showing newer senses of *engag* with*

Deliver

The verb *deliver* (and its verbal inflexions) is a case of a perfectly good literal verb which has been appropriated from management rhetoric by politicians for decades, and recently, to fulfil a rhetorical role in the service of the Coalition government. In its established earlier sense, *deliver* denoted the carrying out and completion of a physical process, especially in the contexts *deliver + a letter/baby/performance/ lecture*. In our data, *deliver* plays a metaphorical role, as illustrated in the extract:

(2) 11/08/10 Cameron delivered his vision for the "big society"

From mid 2011, it shows a growth spurt, collocating with plural nouns with the semantics of public service management: *services*, *improvements*, *results*, *benefits*, *savings*, *growth*, *quality*, *outcomes*; and with mass nouns: *value*, *ability*, *service*. It substitutes for the traditional verb collocates of these nouns, which formerly included *provide*, *supply* or possibly *furnish*. The new contexts are exemplified in Figure 26.

15/07/11	providers contracted to **deliver** some of the **key employment services**
15/07/11	no evidence that the private sector is able to **deliver better outcomes**.
15/07/11	Pooling expertise will help the NHS **deliver a truly excellent service**.
15/07/11	fledgling mutuals are being set up to **deliver public services**
15/07/11	will they simply have to collaborate to **deliver the required services**?
17/07/11	The programme will not continue to **deliver the desired results**

Figure 26. Contexts for the search term *deliver** in 2011

The motivation for this latest semantic manoeuvring seems to be to amplify the rhetoric of the *big society* agenda by stressing the full execution and completion of a service or activity, implying that there is extra commitment and that the task involved is an important one, and thereby making a more positive and dynamic impression on the reader (and the UK citizen under the new moral imperative to shoulder responsibility).

2.7 Grammatical neology

Impact

In the *big society*, there is a particular emphasis on work which delivers *outcomes*, *results*, and *impacts*. Neologisms can emerge in the form of existing words like these with new syntactic patterns or grammatical word classes. A case in point is the noun *impact*, in the verb phrase *have an impact on*, illustrated in Figure 27. With this traditional syntax, it increases considerably in frequency during the period under study.

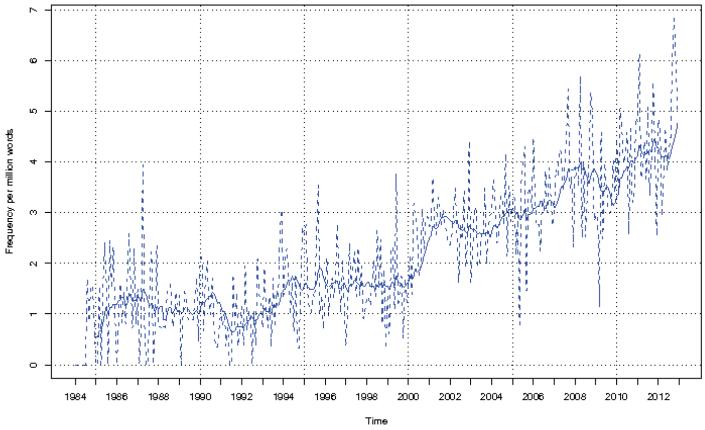


Figure 27. Time graph for [*hav**|*has*|*had*] *an impact on* (3,178 occs.)

Recently, there has also been a sharp increase in the occurrence of *impact on*, as a phrasal verb. This conversion is exemplified in Figure 28.

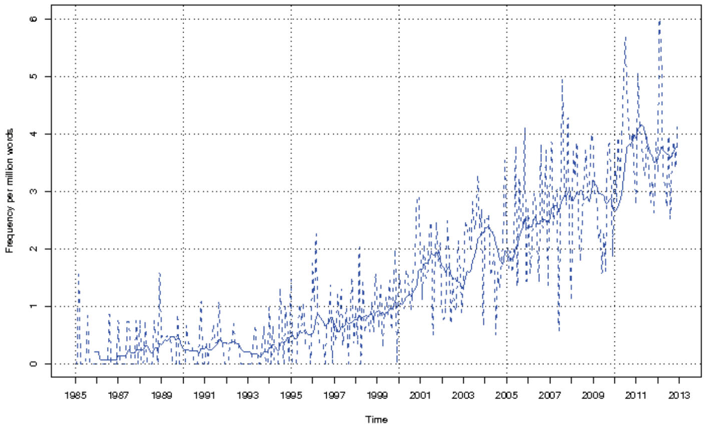


Figure 28. Time graph for the search string *impact**{V*} *on* (2,250 occs.)

A further modification, *impact* + DET(*the*), is beginning to show its face in UK mainstream journalism, and in Figure 29, we see a time-graph for the search pattern *impact**{V*} *the*.

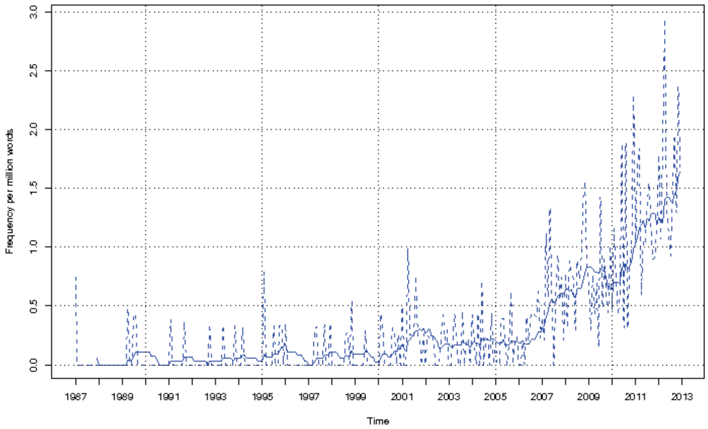


Figure 29. Time graph for the search string *impact**{V*} *the*

Meanwhile, in Figure 30, we see examples with UK provenance dating from late 2011.

07/10/11 This initiative will **positively impact the** livelihood of local farmers
11/10/11 your campaign will not only **impact the UK**, but have ripples through Europe
28/10/11 an encampment on a busy thoroughfare **impacts the** rights of others.
03/11/11 relationships that will **impact the** way we govern society
15/11/11 the banking system could be **severely impacted**," the IMF has said.
28/11/11 It **impacts the** dynamics of the economy and slows the economy
13/12/11 the black hole which may **impact the** overall success of the MDGs.
22/12/11 the economic challenges have **significantly impacted the** solar industry

Figure 30. Extract of contexts for *impact**{V*} *the* (total 601 occs.)

2.8 Productivity and creativity

The meaning of the terms 'productivity' and 'creativity' has been the subject of debate for years (cf. Bauer 1983: 62 and Bauer 2005 for an overview). Unlike other linguists, we use 'productivity' to refer not only to morphological productivity, but also to lexical productivity, by which we mean the process whereby, like morphemes, certain words under certain circumstances become the stimulus and basis for generating more words; and in particular, novel grammatical and lexical structures.

By ‘creativity’, we mean, put simply, the unconventional creation of new words, typically flouting word-formation rules. Fischer (1998: 17) expresses the distinction for us as follows: “Productivity refers to rule-governed word formation processes ... It differs from … creativity, which is unpredictable and not governed by rules”. Though, as she goes on to say, “[i]n reality … productivity and creativity cannot be strictly kept apart” (1998: 17). Both the productivity and the creativity of a particular word fluctuate in frequency over time, for real-world and linguistic reasons (but also due to authorial, editorial and text-type changes) (Baayen and Renouf 1996: 74ff). Neologisms which are fostered by the media typically become productive and creative during (and immediately subsequent to) a topical event (Renouf 2008: 61-92).

One thus expects to find such coinages emerging in the environs of alpha, beta and other core topical terms in our data. In fact, we find that the three alpha/beta neologisms selected here are not very productive or creative. This could be for reasons both social and linguistic: on the one hand, they may be perceived as dealing with serious issues and thus not played with; they may have negative connotations; or journalists may just not like them; on the other, their morphology and grammar may impose restrictions.

Fischer (1998: 17) makes a general distinction “between coining due to word-formation rules, and analogous coining”. We do find, later in the period, some limited analogous coining based on *squeezed middle*, as in Figure 31:

01/06/11	It suggests there is not a “**squeezed bottom**”
02/02/12	MPs voted through a benefit cut for **the “squeezed” bottom half**
13/04/12	I would focus on the elite “**squeezed top**”
18/04/12	The director describes those around 35 as the “**squeezed middle-ages**…”

Figure 31. Contexts for coinages formed by analogy with *squeezed middle*

For the noun phrase *big society*, meanwhile, we find mainly basic coining by word-formation rules, as shown in Figure 32:

19/07/10	the output of the One Nation group contains something **big society-ish**.
04/07/11	the **big society-influenced** New Strategic Direction was a vote-loser.
15/07/11	the need to integrate services, **big society-style**, with communities
15/08/11	Cameron argued that Britain needed **big society-led** moral renewal.
25/01/12	business rates would finance more **big society-type** ventures.
10/12/12	the **Not So Big Society** blog

Figure 32. Contexts for coinages based on *big society*

The function of the coining in Figure 32 is not to name. Its function is primarily located on the “autonomous plane”, where “the organization and maintenance of text structure is the focus” (Sinclair 2004: 53), and where it facilitates the rephrasing of syntax, and thus of information. Its function in collocation with its base form, meanwhile, is deictic (Hohenhaus 2008: 19): typically to convey new information about a given concept. Examples of this discourse function can be seen further in Figure 33. The extent of the length of separation between a node word and possible collocates may be specified in the search string by an asterisk preceding a number. For the revived word *sustainability*, a distance of 6 words (*6) was set for any inflection in the string: *sustainable*/*sustainably*/*sustain*/*sustains*/*sustained*/*sustaining*.

15/10/10	The crucial argument for **sustainability** is financial. “Managing a business **sustainably** saves you money on energy
11/11/10	it’s hard to answer the question about **sustainability**. Will it manage to **sustain** those improvements once the funding is withdrawn
08/03/11	**sustainability** means behaving in an economically **sustainable** way that protects the environment
23/05/12	The first is **sustainability**; we need to make **sustainable** living commonplace.

Figure 33. Contexts for *sustainability* and collocates up to 6 spaces away

Another example of this ‘complex repetition’ pattern (Hoey 1991: 55-56) of morphologically related variants is found in Figure 34, for the search string [*engag** *9 *engag**]. This retrieves all instances of the collocation of two word forms beginning *engag* within 9 words of each other (9 being the maximum distance allowable by our system). As shown in Figure 34, the *engag** followed by *engag** word pairing is often used to make a proposition or statement, and follow it up with a related comment.[15]

07/03/11	That’s where the majority of people are **engaged**, and that is where you need to **engage** with them.
23/05/11	to see how we **engage** with communities and how we can facilitate their own **engagement**.
01/11/11	St Paul’s Cathedral has confirmed it is **engaging** with the protesters. The cathedral’s authorities would **engage** “directly and constructively”
18/11/11	How do we get them **engaged** with a huge project which will **engage** them for the rest of their economic lives?
04/10/11	if students don’t feel like they are really **engaging** with their university, and that the university is **engaging** with them, their education will suffer. This kind of **engagement** can be expensive.

Figure 34. Extract of recent contexts for [*engag** *9 *engag**] (144 occs.)

Across an individual text, this repetition pattern, of morphologically related variants of a lexeme, is one of several types which together form the lexical cohesion which maintains the logical progression of the text (Hoey 1991: 52-74).

Creativity has not reared its head perceptively for the neologisms in this paper. No *squeezy muddledom* or *alarm-clock-Britain-itis*, so far. Since the noun phrase *weapons of mass destruction* was immediately creative once uttered by

George W. Bush in 2002, one might surmise that political neologisms do not attract the same creative response as those which are event-driven; but perhaps the multi-word structure of *weapons of mass destruction* offers more potential for creativity by lexical replacement than other terms.

2.9 Collocation of synonyms

It cannot be a surprise that more than one writer of the language responds to the same new situation by creating a new word to encapsulate it (or pulling into service an existing one). We have illustrated this phenomenon of multiple synonym coinage with the semantic pairings (albeit more contrastive than synonymous) of some alpha and beta neologisms earlier. Such neologistic pairs may compete or be complementary. Boussidan, in her study of semantic change, examines the competing French synonyms *mondialisation* and *globalisation* over time, seeing the former exhibit 'connotational drift' or "progressive changes in connotation" (Boussidan 2013: 6) across time, to survive. But our focus in this paper is on the complementary synonyms which emerge to serve jointly in recording aspects of our extended topical event. If these 'compete', it is in their expression of a real-world, ideological competition between politicians at a period in history. As said earlier, we are not concerned about the future survival prospects for these terms beyond our period of study.

Take, for example, the synonymous alternatives *outreach* and *engagement* exemplified in Figure 35. In response to the search command [*outreach* *8 *engagement*], seven co-occurring pairs were retrieved in the period 2010-2011. These occur in varying sense relationships to each other.

(1) 15/01/10 he's also keen on what he calls "**outreach**", which is **engagement** with the public.

(2) 16/12/10 Our **outreach** ranges from one-on-one **engagement** with elites to press interviews to mass-audience

(3) 10/05/11 Kew plans to convert Evolution House into a community **outreach** and **engagement** centre.

(4) 26/10/09 His commitment to **outreach** and **engagement** with social and political issues

(5) 13/05/11 the Committee's support for a programme of **outreach** and public **engagement**.

(6) 20/05/09 a programme of speeches, conferences, **outreach** events, **engagement** with young people and foreign visitors.

(7) 30/09/11 Awlaki was known for his "interfaith **outreach**, civic **engagement**, and tolerance"

Figure 35. Contexts for the search string [*outreach* *8 *engagement*]

In examples (1) and (2), *outreach* is glossed as being a kind of *engagement*. In (3), *outreach* and *engagement* are presented in a coordinated phrase, presumably to indicate the full range of meaning conveyed by the two words, both in their similarities and their slight differences. In (4) and (5), the difference is made explicit, in the modification of *engagement* to specify an aspect of its distinctiveness from *outreach*. In (6) and (7), both words are modified, to differentiate them and stretch their combined range of meaning.

2.10 Collocation of converging words

The new and modified words and uses to which we have referred hitherto begin to combine in texts to contribute to the general tenor and topic of media and public focus. Figure 36 provides an example of two favoured terms: *engage with* and *deliver*. These are core examples of the kind of inflated language increasingly used by mid-2011 to assert that 'important' things are being, or must be, undertaken and achieved (in spite of the current economic recession and funding cuts).

(1)	04/02/10	only a requirement for the venture to **engage with** industry **to deliver** an agreed specification **(= specify sth)** can **achieve** widespread market success **(= be commercially successful)**
(2)	12/03/10	using creativity to **engage with those who use and those who deliver** social innovation **(= innovate)**
(3)	13/07/11	CITB has to **engage with those who plan and deliver** training **(= train)**
(4)	04/11/10	it urged us to **engage with technology, science and people to deliver** the solutions **(= solve problems)**
(5)	07/04/10	Real improvement will mean **engaging with the staff who deliver** local services **(= serve local people)**
(6)	24/09/10	the government **engages with the renewable industry to deliver** the goal of decarbonising the energy sector **(= decarbonise)**
(7)	29/04/11	government has to **engage with** business **to deliver** the transition **(= move)** to a low-carbon economy

Figure 36. Contexts for the search string [*engag* with* *6 *deliver**] (73 occs.)

In each case in Figure 36, a simpler and more concise expression has been sacrificed in the service of bureaucratic rhetoric. Part of this register consists in favouring lengthy de-verbal constructions over the traditional verb. One should not be too reductionist, but a series of shorter, everyday formulations can be envisaged. In example (1), 'to deliver an agreed specification' could more simply be rendered as 'to specify something'. In (2), 'deliver social innovation' could have been 'innovate'. In (3), 'plan and deliver training' could just be 'train'. The use of a process noun – 'to specify', 'to innovate', 'to train' – is strategic, in that it avoids the need to spell out the object: these verbs do require a clear object. But in (4), 'to deliver the solutions' could be 'solve the problems'; in (5), 'deliver local services' could be 'serve local people'; in (6), 'to deliver the goal of decarbonising' could have been 'to decarbonise'; and in (7), 'to deliver the transition' could have been ‘to move’ or ‘to change’ it. On the other hand, the shift from traditional verb to a de-verbal construction does allow the latter to be pre-modified for further specificity – ‘agreed specification’ and ‘social innovation’.

The building up of a register which everybody in public life (social services, politicians, industry) needs to use adds weight to these words. In Figure 37, we see *engagement* and its inflections collocating with *stakeholders*, as well as with other words emblematic of the ideological and political climate.

29/07/09	GNM’s **engagement with stakeholders** is embedded within the organisation
27/07/09	**engaging with stakeholders** and involving them in GNM’s sustainability approach.
14/10/09	companies have been able to promote the idea that they **are engaging with stakeholders**
27/10/10	Its crisis communications strategy made little effort to **engage with stakeholders**
08/03/11	It will give enterprises a better chance of **engaging with stakeholders**
28/04/11	Effective social media is about having an open mindset to **engage with stakeholders**.
13/04/11	a small charity requires people who can **engage with stakeholders**
05/07/11	we will **engage with stakeholders** including the trade-offs.
04/07/11	we intend to **engage with stakeholders** on these issues

Figure 37. Contexts for the search string [*engage* + *stakeholders*]

In our data, we even find combinations of three or more of these converging words: see Figures 38 and 39.

17/07/09	the commission needs to do more in terms of **delivery** and **engagement** with **stakeholders**
14/06/10	The answers can help to **deliver** clearer priorities, better **engagement** with **stakeholders** and improved measurement of **outcomes**

Figure 38. Contexts for [*delivery* + *engagement* + *stakeholders*]

28/05/05	these lyrics enabled the government to increase public understanding of privatisation policies and constituted effective communications **outreach** and **stakeholder engagement**
27/05/05	We supported Tanzania for many years in communications **outreach** and **stakeholder engagement**

Figure 39. Contexts for [*outreach* + *stakeholder* + *engagement*]

The extension of this co-occurrence of two and three combined new words to something approaching a new register is well illustrated by a *Guardian* journalist on 5 July 2011, in an article headed “(Health Secretary) Andrew Lansley’s monkey puzzles”:

> One problem is that Lansley speaks largely in jargon. In Lansley’s world, people are forever “*commencing substantial piloting*”. Nobody ever talks to the people who might be affected; instead they “engage with stakeholders”... The jargon goes on. “We want to undertake engagement in workplace development strategy in relation to care and support ...We need to ensure an engagement with stakeholders and the public, to understand what the public’s attitude would be in circumstances where they have greater clarity about their potential care costs if they are willing to engage with financial services products ...”. That seems to mean, roughly: “We need to tell people how they can insure against their old age”.

3. Conclusion

In this paper, we have studied neologisms which enter the language to express a topical idea or refer to a new concept or event (Hohenhaus 2008: 17). In particular, we have looked at neologisms emerging in the second half of 2011, which portray society as one where the government holds ordinary citizens accountable for keeping society together and for exceeding past achievements, paradoxically just when the national economy is shaky and less funding is available to support jobs and public infrastructure. We have then studied how such lexical neologisms are bolstered in this task by other words. Some are new: formal, semantic and grammatical neologisms; new synonyms; inflectional and derivational variants. Existing words are temporarily diverted and modified to contextualise these new words, often drawn from a specialised domain (managerial, corporate, technical). All these items collocate and form a patchy network of inter-collocating words which, if sustained by media interest, politicians, public institutions, could evolve into a potential new ‘register’.

What our study across a whole news corpus represents is a serendipitous discovery of linguistic facts, some of which it transpires have been investigated in a more targeted way by Normal Fairclough (2000), a founder of critical discourse analysis. Taking a set of political speeches by Tony Blair, he viewed aspects of New Labour rhetoric through the prism of Old Labour texts, thereby gaining the perspective of language change which our diachronic corpus provides. Fairclough’s approach is not strictly corpus linguistic, in that his starting point is conceptual. What is striking in our findings is the similarity of language use in his Labour texts and our evolving Conservative data, in terms of the manipulation of public perception, and the concentration and co-occurrence of loaded words.

Looking back in early 2014 at our chosen news data of 2009-2012, particularly the second half of 2011, we now see that it amounted primarily to management speak, presented for daily public consumption via media reportage by government, public institutions such as education, health and social services, and business. Spiced up with a few emblematic slogans and rhetorical metaphors, it developed a particularly resonant linguistic character. By late 2013, much of this jargon has already passed into disuse or been superseded. The alpha term *big society* continues to feature significantly in our corpus through 2012, but much of that mention is ironic, critical, or referring back to the era when the swiftly discredited *big society* was still in vogue, or to its subsequent unintended consequences.

Big society is ripe for replacement by a catchword for the next election campaign: Ed Miliband, Labour's shadow PM, seeks in 2012 to appropriate the Tories' intermittently recurring *One Nation* mantra,[16] just as Tony Blair, the 'New Labour' Shadow PM, did in his 1997 election speeches.

Notes

[1] 'Language convergence' is used in the field of language contact to refer to change caused by bilingual speakers mutually borrowing morphological and syntactic features, making their typology more similar. Meanwhile, 'linguistic convergence' (Giles et al. 1991) is employed in Communication Accommodation Theory to refer to the tendency in social interaction for an adjustment of vocal patterns and gestures, to accommodate to others.

[2] Project funded by the Science and Engineering Research Council, the Department of Trade and Industry, Collins Publishers, Nimbus Records, BRS Software Ltd. (DTI Project No. IED4/1/2197; EPSRC Grant No. GR/F99496)

[3] Project funded by the Science and Engineering Research Council, the Department of Trade and Industry, the Press Association and FT Profile. (EPSRC Grant No. GR/L08243/01)

[4] A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colours. The term "heatmap" was originally coined by software designer Cormac Kinney in 1991, for a 2D display depicting real time financial market information. This tool was discovered and adapted for linguistic use within the WebCorpLSE project in RDUES by Matt Gee.

[5] This was uttered in her third term in office in 1987. Exact words recorded in an interview with PM Margaret Thatcher in the *Woman's Own* magazine, 23 September 1987.

[6] Mr Reagan speaking at a fund-raising luncheon for a congressman in Louisiana.

[7] The counterpart to *big* is *small*, not meaning 'diminutive in size', but 'reducing the level of responsibility for running the nation by the

party/sector in question'. The Cabinet Office website enumerates three major parts to the Coalition agenda: community empowerment, opening up public services, and social action.

8 David Cameron promoted it in 19 July 2010 with a speech at Liverpool Hope University.

9 The Care Quality Commission, independent regulator of all health and social care services in England.

10 Notably Hammersmith and Fulham.

11 And incidentally earlier references to social cleansing in some Latin American countries, which evolved in meaning from 'dispossession' to 'genocide'.

12 Popularized by British politician Nicholas Ridley, Tory Secretary of State for the Environment, 1987-1989, who himself opposed a low-cost housing development near his own property, and thus earned himself the title of 'NIMBY'.

13 *Imby* and *imbys* also appear the corpus, though not in response to *nimbyism* in its new rhetorical contrast with *localism*.

14 On the web back in 2004, for example, UCL Professor O'Hara is recorded as saying to a Chinese academic audience: "you have to remember who your stakeholder is when you are teaching English".

15 The final example in Figure 34 incidentally garners a third inflection to add to the pair.

16 Coined in 1868 by Tory PM, Benjamin Disraeli.

References

Baayen, R. H. and A. Renouf (1996), 'Chronicling *The Times*: productive lexical innovations in an English newspaper', *Language*, 72: 69-96.

Bauer, L. (1983), *English Word Formation*. Cambridge: CUP.

Bauer, L. (2001), *Morphological Productivity*. Cambridge: CUP.

Bauer, L. (2005), 'Productivity: theories', in: P. Štekauer and R. Lieber (eds) *Handbook of Word-Formation*. Dordrecht: Springer. 315-334.

Biber, D. and E. Finegan (1994), *Sociolinguistic Perspectives on Register*. New York: OUP.

Boussidan, A. (2013), *Dynamics of Semantic Change: Detecting, Analyzing and Modelling Semantic Change in a Corpus of Short Diachrony*. Unpublished PhD thesis, Université de Lyon, L2C2, ISC.

Crystal, D. (1991), *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Fairclough, N. (2000), *New Labour, New Language?* London: Longman.

Ferguson, Ch. A. (1994), 'Dialect, register and genre: working assumptions about conventionalization', in: D. Biber and E. Finegan (eds) *Sociolinguistic Perspectives on Register*. New York: OUP. 15-20.

Fischer, R. (1998), *Lexical Change in Present-day English: A Corpus-Based Study of the Motivation, Institutionalization and Productivity of Creative Neologisms*. Tübingen: Gunter Narr Verlag.

Giles, H. and N. Coupland (1991), *Language: Contexts and Consequences*. Milton Keynes: Open University Press.

Giles, H., N. Coupland and J. Coupland (1991), 'Accommodation theory: communication, context, and consequence', in: H. Giles, J. Coupland and N. Coupland (eds) *Contexts of Accommodation: Developments in Applied Sociolinguistics*. Cambridge: CUP. 1-68.

Halliday, M. A. K. (1964), 'Comparison and translation', in: M. A. K. Halliday, A. McIntosh and P. Strevens, *The Linguistic Sciences and Language Teaching*. London: Longman. 111-134.

Halliday, M. A. K. and R. Hasan (1976), *Cohesion in English*. London: Longman.

Hoey, M. P. (1991), *Patterns of Lexis in Text*. Oxford: OUP.

Hohenhaus, P. (2008), 'How to do (even more) things with nonce words (other than naming)', in: J. Munat (ed.) *Lexical Creativity, Texts and Contexts*. Amsterdam: Benjamins. 15-38.

Kehoe, A. and M. Gee (2009), 'Weaving Web data into a diachronic corpus patchwork', in: A. Renouf and A. Kehoe (eds) *Corpus Linguistics: Refinements and Reassessments*. Amsterdam: Rodopi. 255-279.

Kitson, M., R. Martin and P. Tyler (2011), 'The geographies of austerity', *Cambridge Journal of Regions, Economy and Society*, 4: 289-302.

Marchand, H. (1969), *The Categories and Types of Present-Day English Word-formation*. 2nd edition. Munich: Beck.

Mollet, Eugene, Alison Wray and Tess Fitzpatrick (2011), 'Accessing second-order collocation through lexical co-occurrence networks', in: Herbst, Th., S. Faulhaber and P. Uhrig (eds) *The Phraseological View of Language: A Tribute to John Sinclair*. Berlin: De Gruyter Mouton. 87-122.

Pacey, M., A. Renouf and A. Collier (1998), 'Refining the automatic identification of conceptual relations in large-scale sorpora', in E. Charniak (ed.) *Proceedings of the Sixth Workshop on Very Large Corpora, ACL/COLING, Montreal, 15-16 August 1998*. San Francisco: Morgan Kaufmann Publishers. 76-84.

Renouf, A. (1993), 'A word in time: first findings from dynamic corpus investigation', in: J. Aarts, P. de Haan and N. Oostdijk (eds) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi. 279-288.

Renouf, A. (2008), 'Tracing lexical productivity and creativity in the British media: the Chavs and the Chav-nots?', in: J. Munat (ed.) *Lexical Creativity, Texts and Contexts*. Amsterdam: Benjamins. 61-89.

Renouf, A. (2012a), ‘A finer definition of neology in English: the life-cycle of a word’, in: H. Hasselgård, J. Ebeling and S. O. Ebeling (eds) *Corpus Perspectives on Patterns of Lexis*. Amsterdam: Benjamins. 177-208.

Renouf, A. (2012b), ‘Defining neology to meet the needs of the translator’, *Neologica*, 6: 17-42.

Renouf, A. (forthcoming), ‘Hapax legomena: a study of newly-occurring adverbs in news text’, in: M. J. Lopez-Couso et al. (eds) *Proceedings of 34th ICAME conference, Santiago de Compostela, May 22-26, 2013*.

Renouf, A. & L. Bauer (2001), ‘Contextual clues to word-meaning’, *International Journal of Corpus Linguistics*, 5: 231-258.

Sinclair, J. McH. (2004), ‘Planes of discourse’, in: J. McH. Sinclair, *Trust The Text: Language, Corpus and Discourse*. Ed. with R. Carter. London: Routledge. 51-66.

Strang, B. M. H. (1970), *A History of English*. London: Methuen.

Wardhaugh, R. (1986), *Introduction to Sociolinguistics*. 2nd edition. Oxford: Blackwell.

English *amid(st)* and *among(st)*: a contrastive approach based on Norwegian and Swedish translation

Thomas Egan and Gudrun Rawoens***

*Hedmark University College
**Ghent University

Abstract

This chapter examines the hypothesis that translation equivalents may be employed to cast light on the semantic network of a lexeme in its original language. The lexemes investigated are amid(st) *and* among(st)*, which are commonly taken to overlap in meaning. The data consist of all tokens of both prepositions in the English language original texts that are common to both the English-Norwegian Parallel Corpus and the English-Swedish Parallel Corpus. A particular question addressed is how the various senses of* among(st) *are related to each other* (amid(st) *is always used to code a* SETTING*). All tokens of* among(st) *are first sorted into semantic classes using normal corpus linguistic methods. The translations into Norwegian and Swedish of the various senses are then examined with an emphasis on the similarities and differences between them. The basic hypothesis is that the senses that are translated in similar ways in a particular language are felt to be more closely related by users of that language than senses that are translated in very different ways. The results lend some support for the hypothesis that translation equivalents can be used as a basis for a semantic classification of polysemous lexemes.*[1]

1. Introduction

The increased availability of parallel and translation corpora has led, in recent years, to something of an explosion in the area of corpus-based contrastive studies. Given this general increase in interest in such corpora on the part of researchers, it is perhaps surprising that they have not formed the basis for more contrastive work on prepositions, which are often taken to be the most intractable of parts of speech, causing innumerable problems for foreign and second language learners. Some exceptions are Schmied (1998), Paulussen (1999), Garretson (2004), Cosme and Gilquin (2008) and Egan (2012). With the exception of Paulussen's (1999) dissertation, these studies of prepositions have been based on parallel texts and translations between two languages. The data in Paulussen (1999), in contrast, comprise translations between three languages: English, French and Dutch.

The present study also makes use of data from three languages, to wit English original data and Norwegian and Swedish translation data. larger study covering various codings of the semantic notion of betweenness (Egan and Rawoens 2013). More specifically, the data for our study comprise two English prepositions, *amid(st)* and *among(st)*, and their translations into two lang-uages, Norwegian and Swedish. According to Lindstromberg (2010: 89), these two prepositions “have quite specialized meanings which are mostly applied metaphorically”. However, when it comes to differences in meaning between the two, Lindstromberg offers little in the way of explanation. Indeed in a section entitled “*AMID(ST)* VS *AMONG(ST)* & *IN THE MIDST OF*, *IN THE MIDDLE (OF)*” (2010: 94) he makes no mention whatsoever of “among(st)”, despite the promise contained in the heading. Moreover, while Quirk et al. (1985) distinguish the meanings of *between* and *among*, pointing out that the latter relates to what they term “nondiscrete objects”, they merely state of *amid(st)* that it “like *among*, can apply to an indefinite number of entities” (Quirk et al. 1985: 680).

In general, the great advantage of linguistic studies based on translation corpora is the fact that these corpora reveal which lexemes or constructions in language *a* are felt by competent users of both languages to correspond most closely to a given lexeme or construction in language *b* (e.g. Dyvik 1998, 2004, Noël 2003, Garretson 2004, Johansson 2007). In our study we exploit the competence of these language users to shed light on the structure of the English forms as these are refracted through the prisms of both Norwegian and Swedish. The reason our study is based on translations into just these two languages is the existence of corpora, described in Section 2, that contain translations of the same set of texts into both languages.

The chapter is structured as follows. In Section 2 we present our aims, our data and the methodology employed. Section 3 describes first the semantics of *amid(st)* and then gives details of the translation equivalents employed by the Norwegian and Swedish translators respectively. Section 4 follows a similar procedure for *among(st)*. Finally, Section 5 contains a summary and some conclusions.

2. Aims, data and methodology

One of the aims of this study, as originally conceived, was to tease apart the various senses of *amid(st)* and *among(st)*. As it transpired, all the tokens of *amid(st)* in our data code a single sense, a SETTING, as will be seen in Section 3. *Among(st)*, however, is used to encode a variety of predication types. In Section 4 we concentrate on the distinctive features of the preposition on the one hand, and on the similarities and differences between the Norwegian and Swedish translations on the other.

Our basic hypothesis is that the senses of a lexeme, in this case a preposition, which are usually translated by one and the same lexeme (or construction) are likely to be more closely related within the semantic network of the original lexeme than those translated by different lexemes. This is in line with Garretson’s contention that “… if we take as our default assumption that similar forms will be used to translate similar meanings, we must expect that related meanings will be expressed with the same form more often than unrelated meanings will" (Garretson 2004: 23).

The data used are taken from the following two corpora: the English-Norwegian Parallel Corpus (ENPC) (Johansson 1998, Johansson et al. 2002) and the English-Swedish Parallel Corpus (ESPC) (Aijmer et al. 1996: 79–80). Both of these corpora contain fiction and non-fiction texts (extracts of roughly 10,000 words taken from various works). Given the fact that the two corpora cannot be searched at the same time, we first extracted the sentences containing *among(st)* and *amid(st)* from the source texts in each of the corpora separately.[2] Since not all source texts in the ENPC and the ESPC are identical, we then decided to use the English originals common to both corpora and their respective translations only – the extent of the overlap is roughly half the corpus (see also Hasselgård 2007 on the fiction part). Each author analysed the English original texts and classified all tokens of *amid(st)* and *among(st)* independently before comparing classifications and discussing tokens which we had analysed differently, with a view to arriving at a consensus.

The corpus search yielded us a set of 186 tokens in total, 16 for *amid(st)* and 170 for *among(st)*. We classified all the original English tokens of *amid(st)* and *among(st)* in the corpus data in terms of the semantics of the prepositional expressions. For instance what cognitive linguists refer to as the 'landmark' (Langacker 1987: 216, Lindstromberg 2010: 6) of the preposition may code a SETTING, as in 'the cat among the pigeons', or it may code an AGENT, as in 'the discussion among the partners'. We should point out that whereas Lindstromberg employs the term 'Subject' for the head of the phrase containing a prepositional postmodifier, we will stick to 'Trajector', which is more common in the cognitive literature (see, for instance, Langacker 1991: 5).

AGENTS, like EXPERIENCERS and THEMES code participants in a process. Such participants may be coded as subjects or objects in clauses paraphrasing the predication containing the prepositional phrase. Thus the phrase 'the discussion among the partners' entails the clause 'the partners discuss(ed)', in which 'the partners' is the agentive subject. The two other types of participant, EXPERIENCERS and THEMES, will be introduced in Section 4. A fourth type of landmark does not imply a processual relationship but rather a stative one, with the landmark corresponding to the predicative in a copular construction, as in 'Among world leaders, Obama stands out as…', which entails the clause 'Obama is a world leader'. We have used the term PROPER INCLUSION for this sort of landmark. As for tokens with SETTING landmarks, like 'the cat (was) among the pigeons', these do not code a participant in a predication but rather the circumstances in which the proposition in question holds. These types of constituents were labelled 'circonstants' by Tesnière (1959) as opposed to the more central 'actants' and as fulfilling 'circumstantial roles' by Halliday (1970) as opposed to the more central 'participant roles' (see Matthews 1981: 123). SETTINGS do not belong to the core of a proposition. In the words of Radden and Dirven:

> The nucleus of a sentence is set off against the setting and, just like the conceptual core, is based on an all-pervasive figure/ground configuration. This means that the notion of setting refers to the background against which a situation is set. Setting elements provide information such as where and when the event happened, why it happened, the conditions under which it happened, etc. To provide this type of information, the speaker uses lexical resources, which specify the factors surrounding a situation in more detail. (Radden and Dirven 2007: 50).

The speaker may also use grammatical (or lexico-grammatical) resources to code SETTINGS. This is the case in our study, where they are coded by *amid(st)* and *among(st)* phrases.

Next, for the classification of the Norwegian and Swedish translations, we began by adopting Johansson's model (Johansson 2007: 24) in which translations are distinguished according to whether they resemble the originals syntactically. Translations which mirror the syntax of the original are labelled "congruent", whereas translations which differ syntactically are labelled "divergent". The difference may take the form of a paraphrase, for example. We further subdivided the congruent translations according to whether they employ the most frequently used preposition (i.e. Norwegian *blant* and Swedish *bland* for English *among*, for example) or an alternative preposition or combination of particle and preposition. A small number of tokens were not translated into one or other language, or not translated into either of them. These tokens were listed separately.

In the case of *among(st)*, discussed in Section 4, statistical calculations were employed to establish whether the forms of translation of the various semantic classes into Norwegian and Swedish differ significantly from those of the other classes. Our calculations were based on our three main translation categories, translation by the default preposition, by an alternative proposition or by a syntactically divergent form. We employed the Fisher Exact Test with two degrees of freedom for all calculations, since some of our raw numbers were smaller than five. We then compared the results of our two sets of calculations. The degree of (dis)similarity between them may be taken as a measure of support for the basic hypothesis that translation equivalents may be of use in sketching the semantic network of polysemous lexemes.

3. The semantics of *amid(st)*

There are, in all, 15 phrases containing *amid* and a single phrase containing *amidst* in our data. These all denote a SETTING, either spatial (7 tokens), as in (1) or circumstantial (9 tokens), as in (2). We restrict the term 'spatial' to landmarks that are both concrete and static. While a handshake, as in (2), may be concrete in the sense that one can feel the pressure exerted by the hand, it is necessarily ephemeral. Nevertheless the landmark still codes a SETTING, as defined in the previous section, since it codes the background against which the leaving of Celia takes place.

(1) It was next to a corner site **amid** other dwellings of similar restrained elegance [….]. (JH1)

(2) As the meeting broke up, Celia left first, **amid** smiles and friendly handshakes. (AH1)

Etymologically, *amid* has evolved from the complex Old English preposition *on middan*, meaning 'in the middle of'. The earliest examples of the prepositional construction in the OED contain a complement in the genitive or dative and all code a SETTING, either spatial or circumstantial, as indeed do all the later cited examples.

When we look at the translations of *amid(st)* into the two target languages, we find a variety of forms used, an overview of which is presented in Figure 1. The possible translations into Norwegian and Swedish are categorised using the model described earlier with congruent and divergent translations.

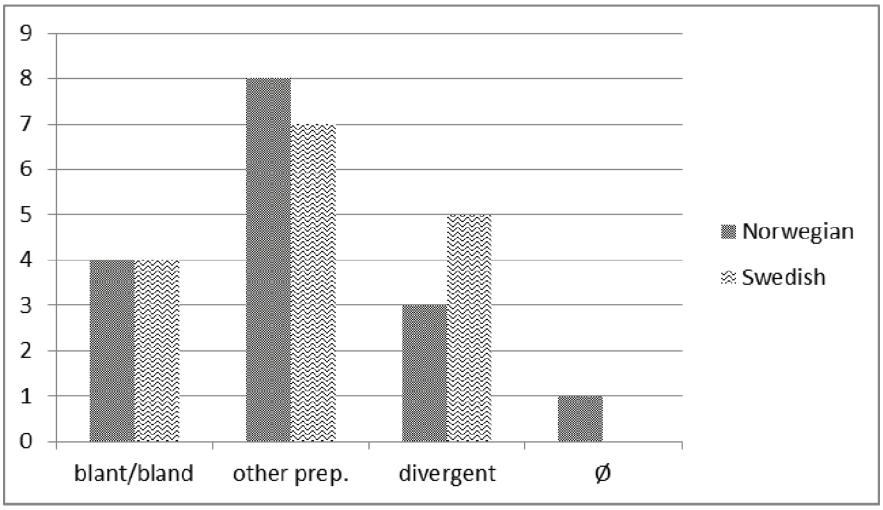


Figure 1. *Amid(st)* translated into Norwegian and Swedish

In the group of congruent translations, the most frequent single translation equivalents in our corpus in Norwegian and Swedish are *blant* and *bland* respectively. Both of these prepositions are descended from Old Norse *bland* (related to the verb *blandan*, meaning 'mix, combine') which was used in the prepositional phrase *i bland* (lit. 'in mixture with', also used in the sense 'have sex with') to mean 'among/together with'. Although no longer used in present-day English, the phrase *i bland* was borrowed by Middle English as *in bland* as in the examples (3) and (4) from the OED. The sentence in (3) is a description of a woman's cheeks.

(3) Boþe quit and red **in-blande**. (c 1340 Gaw. and Gr. Knt. 1205)
'Both white and red intermixed'

(4) In batail […] **in-bland** with þe Grekis. (a 1400 Alexander (Stev.) 2786)
'In battle […] together with the Greeks'

It should be noted that the group of congruent translations containing both Norwegian *blant* and Swedish *bland* in our translation data is relatively small. There are in fact only three instances, one of which is cited as (5). In addition, one occurrence of *blant* in the Norwegian translations corresponds to a divergent translation in Swedish (in which the predication is encoded in a temporal clause). In another case we find *bland* in Swedish and the preposition *under* (= under/during) in Norwegian.

(5) It was next to a corner site **amid** other dwellings of similar restrained elegance ….. (JH1)
Norw.: Det lå vegg i vegg med et hjørnehus, **blant** andre boliger av samme beherskede eleganse
Swed.: Huset låg närmast en hörntomt **bland** andra bostadshus av lika behärskad elegans

As is obvious from Figure 1, both Norwegian and Swedish translators prefer to employ other prepositions than *blant*/*bland*, or combinations of particles and prepositions. These represent roughly half of the instances in the translation data: eight for Norwegian, seven for Swedish. The prepositions and combinations of particles and prepositions employed are listed in Table 1.

Table 1. Congruent translations equivalents with prepositions other than *blant*/*bland*

	Norwegian	Tokens	Swedish	Tokens
Congruent: other prep.	*midt i*	2	*under*	2
	midt under	2	*efter*	1
	etter	1	*i*	1
	i	1	*mitt i*	1
	under	1	*mot*	1
	midt oppe i	1	*trots*	1

The Norwegian data contain four tokens of compounds containing *midt* (= middle), which is cognate with *amid*, two of *midt i* (= in), as in (6), one of *midt oppe i* (= up in), as in (7), and one of *midt under* (= during). As for the remaining four tokens in this group, we find the prepositions *etter* (= after), *i* (= in) and *under* (= under).

(6) **Amid** the pressures of their professional lives, Andrew and Celia found time to look at houses for sale. (AH1)
Norw.: **Midt i** all travelheten (= In the middle of) […]
Swed.: **Trots** yrkeslivets påfrestningar (= Despite..) […]

In the Swedish translations we find only one token where the preposition *mitt* is used, viz. *mitt i*, as illustrated in (7). The other prepositions used in the translations are *efter* (= after), *i* (= in), *trots* (= in spite of), *under* (= under) and *mot* (= against).

(7) …**amid** the birthday foliage of a high-backed seat. (JC1)
Norw.: **midt oppe i** bladverket til en høyrygget fødselsdagsstol
Swed.: **mitt i** den festliga grönska som prydde en stol med högt ryggstöd.

A final point worth noting is that both Norwegian and Swedish contain a number of divergent translations, three and five respectively. Most of these contain a form meaning something like ‘surrounded by’. In (8), for instance, we find a divergent translation in the Norwegian *omgitt av* and in the Swedish *omvärvd av* in (9).

(8) The first waves of landing craft, packed with sodden men, bucketed towards the beaches through the surf **amid** the rippling flashes and explosions of fire […] (MH1)
Norw.: .. **omgitt av** kaskader av lynglimt og eksplosjoner (=surrounded by)
Swed.: .. **till ackompanjemang av** ett pärlband mynningsflammor (= to the accompaniment of)

(9) .. a hysterical woman in a provocative nightdress, shrieking **amidst** a lot of flames […] (MD1)
Norw.: .. **midt i** flammene
Swed.: .. **omvärvd av** lågor (=surrounded by)

To sum up, the low number of occurrences of *amid(st)* in our material does not allow us to draw any firm conclusions about its semantics on the basis of the translations into Norwegian and Swedish. However, what we can say is that the relatively small number of occurrences among translations of *amid(st)* of *blant*/*bland*, both of which are the most commonly used translation equivalents of *among(st)*, as we shall see in the next section, may point to a difference in the semantics of the two English prepositions in our study.

4. The semantics of *among(st)*

There does not appear to be any semantic difference between the two forms *among* and *amongst*. Indeed, the OED does not contain any separate definitions of *amongst*, referring the reader to the definitions of the sub-senses of *among*. For our own part, we did not encounter any semantic differences in our material, the difference being primarily stylistic (according to Lindstromberg 2010: 93, "*AMONGST* is a slightly more literary variant"). For this reason we decided not to distinguish between the two forms in our analysis.

Whereas all tokens of *amid(st)* in our data encode SETTING (cf. Section 3), *among(st)* is used to encode a variety of predication types. Etymologically, *among(st)* has evolved from the complex Old English preposition 'on ʒemang', meaning 'in the company of'. The earliest examples of the prepositional construction in the OED contain a complement in the genitive or dative and all code a SETTING, either spatial or circumstantial. In Middle English the phrase came to be used to code other sorts of relationships, including a now obsolete temporal one. We distinguish six predication types in our material according to the thematic role coded by the landmark of the preposition. The most common of these, with 95 tokens (out of a total of 170 tokens), is the SETTING sense, either spatial or circumstantial, illustrated here by (10) and (11) respectively.

(10) They wander **amongst** the fruits of the earth and sea. (BO1)
(11) She felt guilty about missing church that day, but if God were everywhere, surely He was here **among** so much natural beauty and peace. (GN1)

The second most common sense, instantiated by 26 tokens, we labelled THEME, by which we mean a participant that is "affected by an action or neutrally involved in a situation" (Radden and Dirven 2007: 269). According to this definition PATIENTS, labelled 'incremental themes' by Dowty (1991: 567), comprise a subset of THEMES. The distinction between more and less PATIENT-like THEMES is not germane in the context of our study, as all our THEMES involve participants who are 'neutrally involved' in the words of Radden and Dirven. Thus in (12) the rural under 25's are said to be unemployed, without there being any indication of who may have made them so.

(12) Unemployment **amongst** the rural under 25's is reckoned, currently, at around 60 per cent. (FW1)

There are 32 tokens in which the landmark of the preposition codes the trajector of a process denoted by the trajector of the preposition, in a manner similar to a subjective genitive. These may be divided into tokens where the landmark codes an AGENT, as in (13) in which it is the secretaries who are engaged in gossiping, and tokens where it codes an EXPERIENCER, as in (14), in which it is the painters who feel envy.

(13) But there the gossip **amongst** the secretaries and clerks was way off mark. (JC1)

(14) But I only make enough to generate envy, **among** other painters, not enough so I can tell everyone else to stuff it. (MA1)

Nine tokens of the AGENT category are reflexive. In example (15), the function of the reflexive prepositional phrase is to limit the process coded by the verb, in this case talking, to a restricted number of participants, commensurate with the trajector of the process.

(15) What old men want is peace and informality, and the chance to talk **amongst** themselves like smutty boys. (JC1)

There is one relatively common type of predication, represented by 15 tokens, which contains a copular verb. In these the landmark of the preposition codes a category to which the trajector belongs. As explained in Section 2, we have used the term PROPER INCLUSION for these tokens (compare 'category inclusion' in Radden and Dirven 2007: 273). One of these is cited as (16).

(16) Kate and Peter must have been **amongst** the few children who had to plead with their parents to be allowed to attend prayers and assembly and scripture lessons. (MD1)

Finally, there were two tokens which did not fit comfortably into any of these five categories, although they shared characteristics with several of them. In (17) there are various possible permutations of the distribution of the six bathrooms across the three apartments. We have labelled this usage DISTRIBUTION.

(17) When we cleared out his stuff he had three apartments, with six bathrooms **among** them, and every bathtub was piled high with bundles of pictures and sketches and books and manuscripts and whatnot. (RDA1)

Table 2 shows the number of tokens per semantic category (predication type) of *among(st)*, in descending order of frequency.

Table 2. Tokens of *among(st)* per semantic category

Predication type	Number of tokens
SETTING	95
THEME	26
AGENT	22
PROPER INCLUSION	15
EXPERIENCER	10
DISTRIBUTION	2
Total	170

4.1 Norwegian and Swedish translations of *among(st)*

We turn now to the Norwegian and Swedish translations of *among(st)*.[3] Figure 2 contains details of these in terms of the four options we mentioned in Section 2: i.e. translation by the default prepositions *bland*/*blant*, congruent translations containing another preposition, translations containing a divergent syntactic form, and non-translation of the predication.

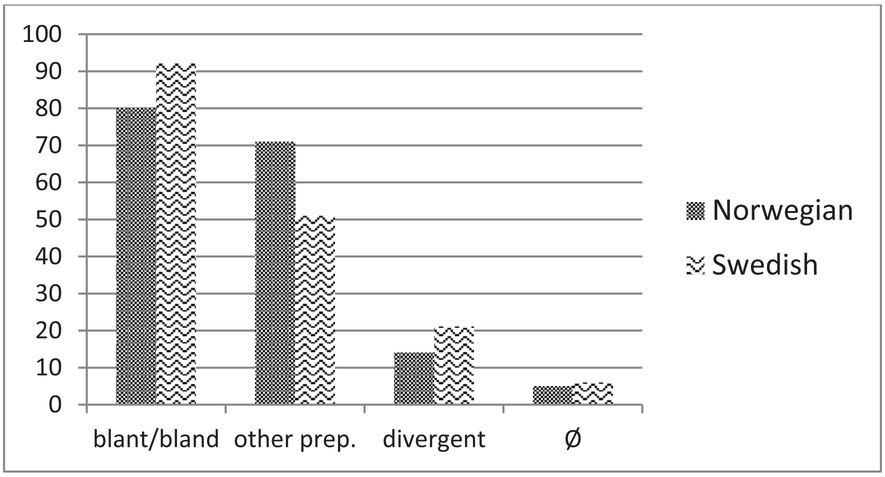


Figure 2. *Among(st)* translated into Norwegian and Swedish

We see in Figure 2 that both Norwegian *blant* and Swedish *bland* are the single most common translation equivalents utilised by the two sets of translators. This is of course why we have chosen to single them out in our analysis. Swedish *bland* is more common than Norwegian *blant*, which is represented by a dozen fewer tokens. Norwegian, on the other hand, has some 20 more tokens which contain an alternative preposition, while Swedish has more divergent tokens. Finally there are only a handful of originals that the translators into both languages have opted not to translate.[4]

Figure 2 may, however, give a false impression of the degree of overlap between the options chosen by the two sets of translators. Indeed, so great appears to be the overlap that one might question one of the premises for our study, which is that we are engaged in comparing two different sets of translations. Perhaps Norwegian and Swedish are not dissimilar enough to provide fruitful data for comparison? Or perhaps some translations into language *a* are based not so much on the original texts as on prior translations into language *b*? However Figure 2 is deceptive in this respect. The actual extent of the overlap may be seen in Table 3, which shows how often the same sort of option is employed for one and the same original by two translators. This information is displayed in graphic form in Figure 3, with the Norwegian tokens on the x axis and the Swedish tokens on the y axis. The small handful of non-translated tokens are mentioned in Table 3 but have been omitted from the figure, since our main interest here lies in comparing forms that are actually used. In any case the numbers involved are so small as would render their representation in the figure indecipherable.

Table 3. Correspondences between Norwegian and Swedish translations of individual tokens

	Norw. *blant*	Norw. other prep.	Norw. divergent	Ø	Total
Swed. *bland*	57	28	5	2	92
Swed. other prep.	15	34	2	0	51
Swed. divergent	6	6	6	3	21
Ø	2	3	1	-	6
Total	80	71	14	5	170

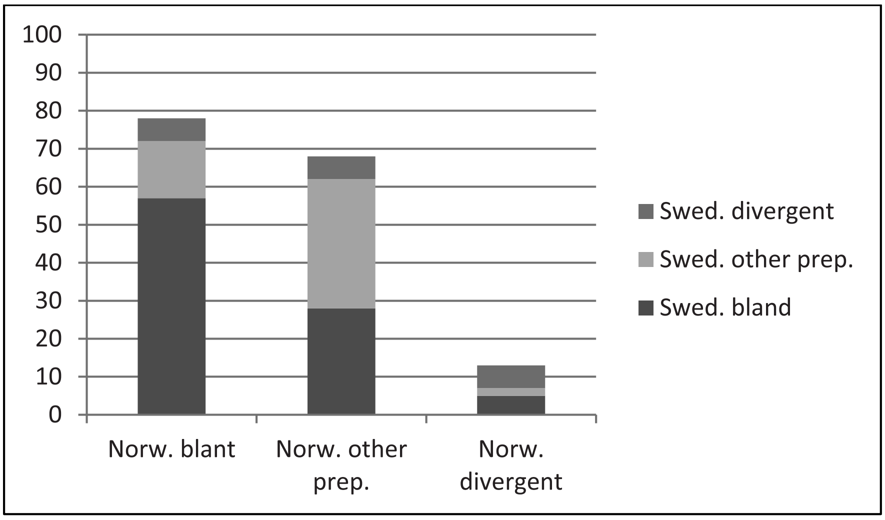


Figure 3. Correspondences between Norwegian and Swedish translations of individual tokens

One can see from a glance at Table 3 and Figure 3 that while the majority of tokens of Norwegian *blant* correspond to Swedish *bland* (i.e. 57 instances), there is a sizable minority of 23 tokens where this is not the case. As for tokens containing other prepositions, only half of these in Norwegian (i.e. 34 out of 71) correspond to prepositions other than *bland* in Swedish. On the evidence of Figure 3, we can safely conclude that we are dealing with two different sets of translations, albeit into quite similar languages.

4.2 Translations of SETTINGS coded by *among(st)*

We now turn our attention to translations of the various types of predication coded by *among(st)*. Of 92 tokens encoding a SETTING that are translated into both languages, 33 (36%) are translated by both Norwegian *blant* and Swedish *bland* as in (18) and (19).

(18) A dying fly buzzed its last song up on the ceiling, **among** the net of cobwebs. (BO1)
Norw.: … **blant** spindelvevene …
Swed.: … **bland** härvan av spindelnät ..

(19) I was obviously going to have a musical time **among** the radiators and stopcocks. (PM1)
Norw.: … **blant** radiatorer og kraner.
Swed.: … **bland** element och flottörer.

40 SETTING tokens are translated into Norwegian and 33 into Swedish by prepositions other than *blant*/*bland*. Of these 21 are translated in this way into both languages and, of these 21 instances, 13 are translated by cognate prepositions in Norwegian and Swedish, such as *i* (= in) in (20) and *mellom*/*mellan* (= between) in (21).

(20) It was in the East End, down **among** the dockland, that he had first started ... (FF1)
Norw.: … nede **i** havnestrøket
Swed.: … nere **i** hamnkvarteren

(21) Sometimes fallow deer can be seen **among** the trees. (RR1)
Norw.: … **mellom** trærne
Swed.: … **mellan** träden

Example (20) is unusual in English in so far as the landmark of *among* is both concrete and singular. Moreover, it is not modified by a plural noun, as is *net* in (18). Mass nouns, both abstract and less often concrete, are often encountered as landmarks of *among(st)*, as in (11). However (20) is the only token in our material with an unmodified singular concrete noun. This usage may well be idiolectal. In any case both the Norwegian and Swedish nouns corresponding to 'Dockland(s)' in English are singular, and both translators have chosen to interpret these as containers within the boundaries of which the activity denoted by the verb is situated. In the case of (21) both translators choose not to employ *blant*/*bland*, the prepositions corresponding most closely to *among*, which would preserve the original's emphasis on the location of the deer, but rather to use the equivalent of *between*, which emphasizes rather the path of perception, with the result that the translations may be paraphrased as ‘if you look between the trees, you will see the deer’.

4.3 SETTINGS coded by *amid(st)* and *among(st)* compared

We saw in Section 3 that *amid(st)* is used exclusively to encode SETTINGS, both spatial and circumstantial. Since *among(st)* is also commonly used to code SETTINGS, the question arises as to the similarities and differences between the two prepositions. One difference is that while *amid(st)* is almost equally likely to code a spatial as a circumstantial SETTING, the latter type is only half as common as the former in the case of *among(st)*. This difference between the prepositions is not, however, statistically significant. Another possible difference, mentioned in the literature, relates to the number of objects which make up the landmark of the preposition. Thus Hickmann and Robert (2006: 4) write that the landmark of *among* consists of ‘several objects’, the landmark of *amid* of ‘numerous objects’. This statement seems to imply that the greater the number of objects comprising the SETTING, the more likely the language user is to employ *amid*. We return to this contention below, but first we compare the translation equivalents of both prepositions in Norwegian and Swedish. These are given in Figure 4, which con contains percentages rather than raw numbers, to better enable comparison between the two, given that *among(st)* is six times more common than *amid(st)*.

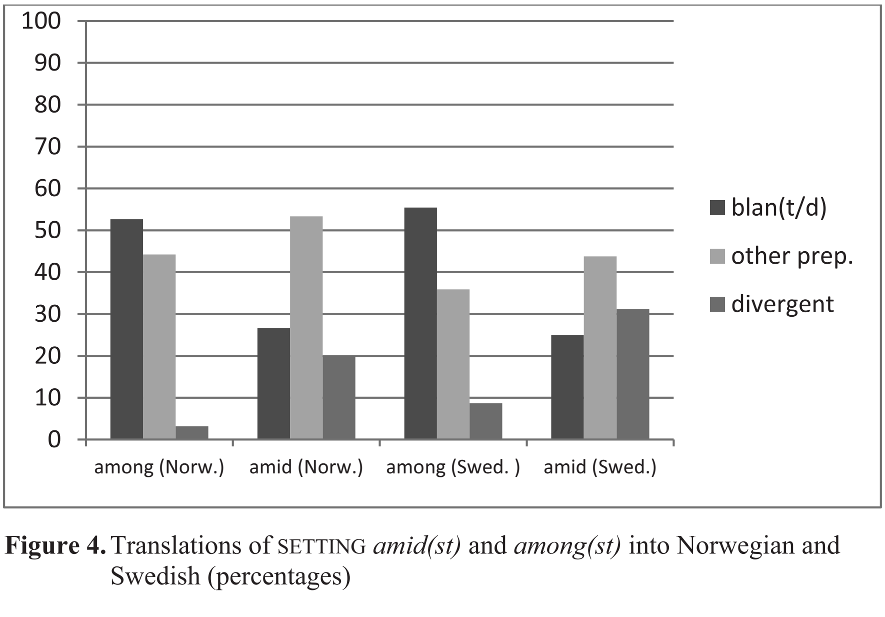


Figure 4. Translations of SETTING *amid(st)* and *among(st)* into Norwegian and Swedish (percentages)

The evidence presented in Figure 4 shows that the *among(st)* SETTING tokens follow the pattern for *among(st)* as a whole, shown in Figure 2, in so far as *blan(t/d)* is the most popular form in both languages, followed by other prepositions and finally divergent constructions.[5] Furthermore, if we distinguish between spatial and circumstantial settings, we may further note that in the case of *among(st)*, approximately 50% of both types are translated by *blant* in Norwegian and *bland* in Swedish. This is in marked contrast with *amid(st)*, where just one of the nine circumstantial SETTINGS is translated by by *blant* in Norwegian and none whatsoever by *bland* in Swedish. If we restrict our attention to the spatial SETTING tokens and try to distinguish between those whose landmarks consist of a restricted number of items and those containing an unrestricted number, we find that in all tokens of *amid(st)* the landmark consists of 'numerous objects', to borrow the term used by Hickmann and Robert (2006). Examples (5) and (8) are typical in this respect. On the other hand, while the landmark of *among(st)* may also consist of an unrestricted number of objects, as in (10), there are also quite a few examples, such as (22) and (23) where the number of items is bounded, though the exact limits are not given by the context.

(22) He kissed the tips of his fingers, speckling his beard with white, and poked **among** the papers on his table, raising puffs from every pile. (PM1)
Norw.:… **igjennom** papirene på bordet (= through the papers)
Swed.:… **bland** papperen på bordet

(23) I sit by myself in the back of the car, **among** the suitcases and the cardboard boxes of food and the coats…. (MA1)
Norw.:… **blant** koffertene og…
Swed.:… **bland** kappsäckarna och...

Since both *amid(st)* and *among(st)* are used to encode SETTINGS, one may ask whether there is any difference in the types of situation that they typically frame. Figure 5 contains percentage details of how often they are used to frame Vendler's (1967) four situation types.

As shown in Figure 5, *amid(st)*, though it can be used to code the SETTING of all four types of predication, is most likely to be used with Activities, as in (6) and (8). It typically encodes the background for some ongoing situation.[6] *Among(st)*, on the other hand, is more likely to occur with a State, as in (10) and (24). And when it is used to locate a State, it is more likely to be translated into Norwegian by *blant* (70%) and Swedish by *bland* (65%) than the other predication types.

(24) Whole families stayed out in the night, huddled **amongst** the ragged ends of their clothes and mattresses. (BO1)
Norw.: … **blant** fillete rester av klær
Swed.: … **bland** sina trasiga stycken av kläder

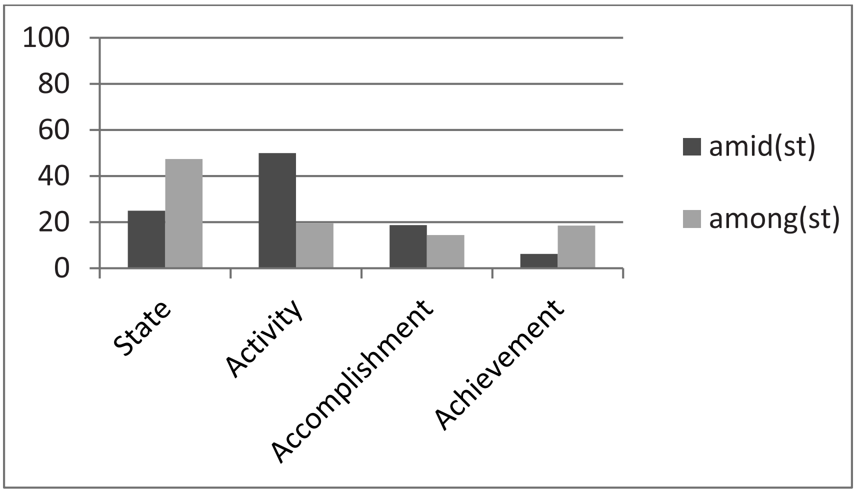


Figure 5. Percentage figures for types of situation the SETTING for which is coded by *amid(st)* and *among(st)*

4.4 Translations of other predication types

Having dealt with tokens coding SETTING, we turn now to the second most frequent type of predication, in which the landmark of *among(st)* codes the THEME of a predication coded by a nominal, in cases where the prepositional phrase fulfils a postmodifying function, as in (12), or a clause where it has more of an adverbial function, as in (25) below. Figure 6 compares translation strategies employed for THEME predications by both sets of translators with those employed for SETTINGS.

(25) Those of us who made such vows were known **among** the Living as abiku, spirit-children. (BO1)
Norw.: **Blant** de levende …
Swed.: … **bland** de levande ...

Figure 6 shows that in both languages there is very little difference between the way SETTING phrases are translated and the way THEME phrases are translated (p=0.447 in Norwegian and 0.885 in Swedish). This indicates that they may be closely related in the semantic network of all three languages. One point worth noting is that translations of THEME phrases exhibit a greater degree of resemblance across languages than do SETTING phrases, with 15 of the 22 phrases translated into both Norwegian and Swedish containing cognate prepositions, 11 *blant*/*bland* as in (26), 2 *av* (=of) as in (27), 1 *med* (=with) and 1 *mellom*/*mellan* (= between).

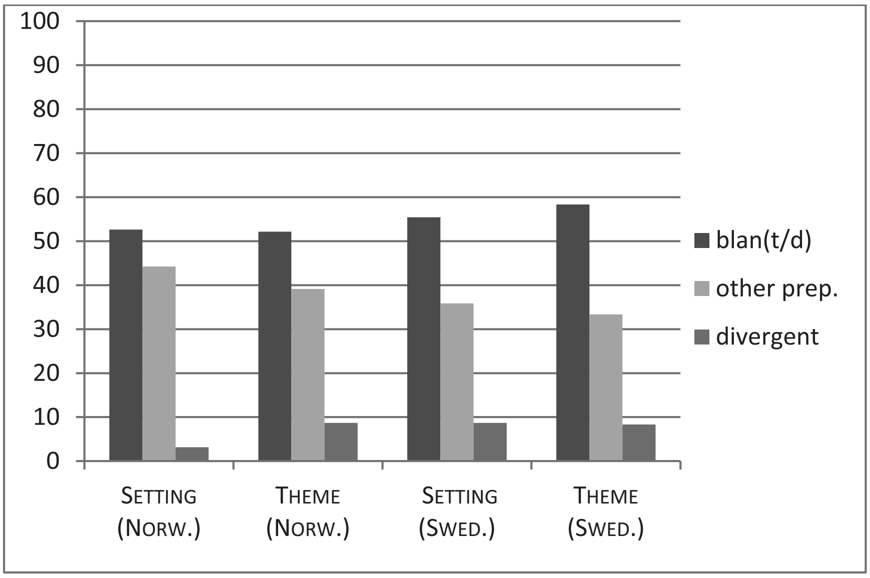


Figure 6. Translations of THEME predications compared to SETTING

(26) **Among** the many things he knew toward the end of his life was that there were many more he did not. (JH1)
Norw.: **Blant** de mange ting
Swed.: **Bland** de många saker
(27) Look, you were one person **among** many whom she knew and talked to ... (MW1)
Norw.: Du var jo en **av** mange
Swed.: Du var en **av** många

Not only is THEME translated in a similar fashion to SETTING in both languages. In addition, there is also no significant difference in either language between the coding of THEME phrases and those of EXPERIENCER, as in (28), PROPER INCLUSION, as in (29), and AGENT, as in (30-32).

(28) There is a growing fear **among** development planners […] (LT1)
Norw.: **Blant** utviklingsplanleggerne
Swed.: **Bland** utvecklingsplanerare
(29) **Among** them was the Rembrandt self-portrait […] (JH1)
Norw.: **Blant** dem var det selvportrettet
Swed.: **Däribland** var Rembrandts självporträtt (= Among them…)
(30) She imported priests of Baal, who quickly acquired a following **among** the northerners […] (KAR1)
Norw.: … **blant** beboerne i nord.
Swed.: … **bland** nordborna.

(31) There were arguments and even brawls every day. **Among** the refugees. (BR1)
Norw.: … **Mellom** flyktningene.
Swed.: … **Bland** flyktingarna.

Reflexive phrases, such as (15) and (32), which we categorised as a sub-set of AGENT phrases, stand out as the only phrase type which is not translated by *blant*/*bland* in either language.

(32) So by a general consensus the party, as it were, metaphorically turned its back on her and talked **among** themselves. (MD1)
Norw.: … samtalte seg **imellom**. (= between themselves)
Swed.: … pratade **med** varandra. (= with one another)

The translations of AGENT phrases not only resemble those of THEME phrases in both languages, they are also similar to those of EXPERIENCER phrases in both. Moreover they also resemble translations of SETTING and PROPER INCLUSION predications in Swedish and, to a lesser extent, in Norwegian.

4.5 Elements of a network for *among(st)*

Having now described how the most common types of *among(st)* predications are translated into Norwegian and Swedish, we proceed to compare the forms used by both sets of translators for the five most common senses in terms of whether they employ the default prepositions *blant* and *bland*, whether they use an alternative preposition, or whether they use an alternative construction. As mentioned in Section 2 we employed the Fisher Exact Test for all our calculations. Our results are presented in Tables 4 and 5 for Norwegian and Swedish respectively.

Table 4. P-values with 2 degrees of freedom for 3 sorts of translation into Norwegian of 5 subtypes of *among(st)*

	EXPERIENCER	PROPER INCLUSION	SETTING	THEME
AGENT	0.885	0.01	0.035	0.394
EXPERIENCER		0.013	0.018	0.205
PROPER INCLUSION			0.002	0.069
SETTING				0.447

Table 5. P-values with 2 degrees of freedom for 3 sorts of translation into Swedish of 5 subtypes of *among(st)*

	Experiencer	proper Inclusion	Setting	Theme
Agent	0.1	0.752	0.039	0.174
Experiencer		0.095	0.16	0.06
proper Inclusion			0.094	0.279
Setting				0.885

If we were to apply a probability level of p=0.05 to the data in the Tables 4 and 5, we would end up with five significant differences in Norwegian and two in Swedish. However, employing such a measure would involve a 40% risk of attributing significance to a non-significant comparison in one of the ten cases in each table. This is because we have in both cases carried out ten calculations on the same data set. If we wish to avoid this risk, we have to adjust the significance level, as pointed out by Gries (in progress) with reference to the network for *through* in Egan (2012). Applying the Bonferroni Correction for multiple tests on the same data set yields a significance level of 0.005 rather than 0.05, which means that the only significant difference we would be left with is that between SETTING and PROPER INCLUSION in translations into Norwegian.

Does this then mean that Tables 4 and 5 cannot be mined for any information at all about the network of senses of *among(st)* in English? We would suggest that, interpreted with care, the p-values in these tables may indicate some plausible cross-linguistic similarities in the construal of the semantic relationships involved in the various sorts of predications. For instance, the mutual p-value for THEME and SETTING is the highest for both of these sorts of landmarks in both sets of translations, indicating a possible closer relationship between a landmark that encodes a neutral role in a predication (THEME) and one that just codes background information (SETTING) than between either of these and the more actively involved AGENT and EXPERIENCER. The latter two, on the other hand, share a very high mutual p-value in Norwegian and the second highest for both in Swedish. The final thematic role, PROPER INCLUSION, patterns most closely with THEME in Norwegian and AGENT in Swedish. It is also the sense that exhibits the greatest difference in p-values between the two sets of translations.

To sum up, the values in the table point to a distinction between two pairs of conceptually linked *among(st)* landmarks, THEME and SETTING on the one hand and AGENT and EXPERIENCER on the other, with PROPER INCLUSION patterning differently in the two sets of translations.

5. Summary and conclusion

In this chapter we have examined all tokens of the two prepositions *amid(st)* and *among(st)* found in the texts that occur in both the ENPC and the ESPC, with a view to investigating whether the translation equivalents in the two languages can shed any light on the semantics of the original items.

We saw in Section 2 that all tokens of *amid(st)* in our material occur in phrases coding a SETTING, either spatial or circumstantial. Perhaps surprisingly, they are most often translated into Norwegian and Swedish by prepositions other than *blant*/*bland*. SETTING is also the most common of six functions coded by *among(st)* phrases. Unlike *amid(st)* tokens, these are most often translated into Norwegian and Swedish by the prepositions *blant*/*bland*. SETTINGS coded by *among(st)* differ from those coded by *amid(st)* in that they are more likely to be spatial rather than circumstantial. They are also more likely to occur with a State, whereas *amid(st)* is more likely to be used for the background of an Activity. At the end of Section 4 we showed how the similarities between translations into both languages may be used to group the various types of landmark complements of *among(st)* into two pairs, THEME and SETTING on the one hand and AGENT and EXPERIENCER on the other. The fifth type of landmark, PROPER INCLUSION, patterns differently in the two sets of translations.

Although much work remains to be done on the semantics of (other) prepositions as reflected in translations into these two and other languages, we would conclude by stating our conviction that the degree of overlap between the network for *among(st)* based on translations into Norwegian and Swedish provides some support for the hypothesis that translation equivalents can shed light on the semantic network of polysemous lexemes.

Notes

[1] We wish to thank two anonymous reviewers for their comments and suggestions. Thanks are also due to Gregory Garretson for his comments on an earlier draft of this paper. In addition Gudrun would like to thank the Swedish Institute for a research grant offered for a research stay at Uppsala University in October 2012 and Thomas would like to thank the research group *Arena for Kultur- og Språkfag* at Hedmark University College for financial support.

[2] The corpora are accessible via an online search tool to holders of an access account. For a description of the works included in the respective corpora see http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf (for ENPC) and http://www.sol.lu.se/engelska/corpus/corpus/espc.html (for ESPC).

[3] Since there are only two tokens of our final category DISTRIBUTION in our material, their translation equivalents cannot be subjected to statistical

analysis. They are therefore not mentioned in the discussion of the translations of the various types of predication. They are however included in the numbers in Figures 2 and 3 and Table 3.

4 One anonymous reviewer points out that in some cases, such as ‘among other things’, translation by *blant*/*bland* is the only available option. This phrase occurs three times in our material. In all three cases it is translated by *bland annat* in Swedish. It is twice translated by *blant annet* in Norwegian and omitted from the third translation. Given the small number of occurrences of such fixed phrases in the corpora we chose not to single them out for separate analysis.

5 One should of course note that SETTING tokens account for over half the total number of tokens in Figure 2, so one is here comparing a subset of data with the whole set.

6 One anonymous reviewer rightly points out that given the small number of tokens of *amid(st)*, 16 in all, there is a danger that one text could skew the proportions completely. Indeed six of the 16 tokens come from the same original text, MH1. Three of these code Activities, two Achievements and one an Accomplishment. The four stative uses occur in four different texts.

Sources

English-Norwegian Parallel Corpus: http://www.hf.uio.no/ilos/english/services/omc/enpc/

English-Swedish Parallel Corpus: http://www.sol.lu.se/engelska/corpus/corpus/espc.html

References

Aijmer, K., B. Altenberg and M. Johansson (1996), ‘Text-based contrastive studies in English. Presentation of a project’, in: K. Aijmer, B. Altenberg and M. Johansson (eds) *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, 4–5 March 1994*. Lund: Lund University Press. 73-85.

Cosme, C. and G. Gilquin (2008), ‘Free and bound propositions in a contrastive perspective: The case of *with* and *avec*’, in: F. Meunier and S. Granger (eds) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: Benjamins. 259-274.

Dowty, D. R. (1991), ‘Thematic proto-roles and argument selection’, *Language*, 67: 547-619.

Dyvik, H. (1998), ‘A translational basis for semantics’, in: S. Johansson and S. Oksefjell (eds) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 51-86.

Dyvik, H. (2004), 'Translations as semantic mirrors: from parallel corpus to wordnet', in: K. Aijmer and B. Altenberg (eds) *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg 22-26 May 2002*. Amsterdam: Rodopi. 313-326.

Egan, T. (2012), '*Through* seen through the looking glass of translation equivalence: a proposed method for determining closeness of word senses', in: S. Hoffman, P. Rayson and G. N. Leech (eds) *Corpus Linguistics: Looking back - Moving forward*. Amsterdam: Rodopi. 41-56.

Egan, T. and G. Rawoens (2013), 'Moving *over* in(to) English and French: A translation-based study of 'overness'', in *Languages in Contrast*, 13: 193-211.

Garretson, G. (2004), *The Meanings of English "of": Uncovering Semantic Distinctions via a Translation Corpus*. Unpublished MA thesis, Boston University.

Gries, S. Th. (in progress), 'Quantitative designs and statistical techniques', in: D. Biber and R. Reppen (eds) *The Cambridge Handbook of Corpus Linguistics*. Cambridge: CUP.

Halliday, M.A.K. (1970), 'Language structure and language function', in: J. Lyons (ed.) *New Horizons in Linguistics*. Harmondsworth: Penguin Books. 140-165.

Hasselgård, H. (2007), 'Using the ENPC and the ESPC as a parallel translation corpus: Adverbs of frequency and usuality', *Nordic Journal of English Studies*, 6.

Hickmann M. and S. Robert (2006), 'Space, language, and cognition: Some new challenges', in M. Hickmann and S. Robert (eds) *Space in Languages: Linguistic Systems and Cognitive Categories*. Amsterdam: Benjamins, 1-15.

Johansson, S. (1998), 'On the role of corpora in cross-linguistic reasearch', in: S. Johansson and S. Oksefjell (eds) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 3-24.

Johansson, S., J. Ebeling and S. Oksefjell (2002), *English-Norwegian Parallel Corpus: Manual*. University of Oslo: Department of British and American Studies. Available online at http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/ENPCmanual.

Johansson, S. (2007), *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: Benjamins.

Langacker, R. (1987), *Foundations of Cognitive Grammar. Volume 1, Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker, R. (1991), *Foundations of Cognitive Grammar. Volume 2, Descriptive Application*. Stanford: Stanford University Press.

Lindstromberg, S. (2010), *English Prepositions Explained.* Revised edition. Amsterdam: Benjamins.

Matthews, P. H. (1981), *Syntax*. Cambridge: CUP.

Noël, D. (2003), ‘Translations as evidence for semantics: an illustration’, *Linguistics*, 41: 757-78.

OED = *Oxford English Dictionary*, available online www.oed.com

Paulussen, H. (1999), A corpus-based contrastive analysis of English *on/up*, Dutch *op* and French *sur* within a cognitive framework. Unpublished PhD thesis, Ghent University.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Radden, G. and R. Dirven (2007), *Cognitive English Grammar*. Amsterdam: Benjamins.

Schmied, J. (1998). ‘Differences and similarities of close cognates: English *with* and German *mit*’, in: S. Johansson and S. Oksefjell (eds) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 255-275.

Tesnière, L. (1959), *Eléments de la syntaxe structurale*. Paris: Klincksieck.

Vendler, Z. (1967), *Linguistics in Philosophy*. Ithaca (N.Y.): Cornell University Press.

Cohesive conjunctions in English and German: systemic contrasts and textual differences

Kerstin Kunz and Ekaterina Lapshinova-Koltunski***

* Heidelberg University, Heidelberg, Germany
** Saarland University, Saarbrücken, Germany

Abstract

The present paper contrasts strategies of cohesive conjunction in English and German system and text. We clarify the notion of cohesive conjunction by discussing conceptualizations in the literature and by comparing cohesive conjunctions to other cohesive strategies. Using theory-informed methodologies we contrast the resources available in the two languages for explicitly establishing conjunctive relations of cohesion. Moreover, we discuss the first findings from our analysis of an English - German corpus of translations and originals, which reveal differences in the textual realizations in terms of frequencies and functions. Our study complements insights about other types of cohesion investigated in the frame of a larger research project.[1]

1. Introduction

The aim of the present study is to contrast strategies of cohesive conjunction in English and German by combining systemic and textual research perspectives with adequate analytical methodologies. *Systemic contrasts* concern the inventory of 1) *cohesive devices* available in each language to establish a conjunctive relation and their structural and syntactic properties, 2) the *logico-semantic relations* they trigger and 3) the structural/ syntactic properties of the textual elements they connect (henceforth termed *connects*). Our discussion of these contrasts is based on theoretical as well as exemplary observations. *Contrasts in the textual realization* of cohesive strategies of conjunction manifest themselves in relative frequencies, which depend on the language and the contexts of situation in which they are realized as well as on the type of production. We therefore draw our findings from a corpus-linguistic analysis of English and German originals and translations in different registers.

Conjunctive relations have been discussed in numerous works, though not always with a focus on cohesive aspects. Some of these approaches incorporate implicit aspects of coherence, most of them are concerned with the systemic resources to establish logico-semantic relations, below and/or beyond the level of the clause or sentence. They either look at phenomena from a rather theoretical perspective and consider examples from multiple languages to construe an abstract model, or focus on the resources of one individual language, e.g. Halliday and Hasan (1976), Martin (1992), Halliday and Matthiessen (2004), Quirk et al. (1985), Pasch et al. (2003), Blühdorn (2008 a/b), Mann and Thompson (1988). Existing contrastive studies of English and German examine textual instantiations of cohesive conjunctions, yet they are either limited to the investigation of individual devices in particular registers, e.g. Becher et al. (2009), Kranich et al. (2011), House (2011), or do not provide empirical evidence, e.g. Fabricius-Hansen (1999) or Doherty (2004). Some studies apply corpus-based processing mechanisms for the empirical analysis of larger sets of conjunctions. For instance, Hutchinson (2005) describes automatic acquisition of knowledge about discourse connectives based on their distributions in texts and focuses on their semantic properties. Corpora are also used in the studies of Lüngen et al. (2006), Stede (2002), Stede (2008a/b) and Dipper and Stede (2006), who apply lexical resources for conjunction extraction. Bestgen et al. (2006) employ corpora to automatically determine the semantics of connectives. Marcu (2000) uses surface-based and statistical methods to identify elementary discourse units and to predict coherence relations between adjacent segments. These studies mainly concentrate on the description of one language only and very few consider register as a variable.

To our knowledge, multilingual corpus-based approaches to the analysis of cohesive conjunctive relations are still missing. Presumably, one main reason is that a comprehensive investigation of this kind cannot be done via fully automatic extraction procedures. It has to integrate different aspects of cohesion such as the resources available in each language under investigation, their lexicogrammatical as well as functional properties and the semantic/ conceptual relations prevalent in different registers. Such a corpus-linguistic analysis requires complex annotations of corpora on multiple linguistic levels in order to yield significant data.

We aim at modeling variation in the linguistic creation of cohesive conjunctions as variables depending on language (English vs. German) and register (written and spoken registers available in the corpus). The present study represents an initial step towards reaching this long-term goal. It is part of a broader contrastive research project on cohesion that additionally looks into cohesive reference, substitution, ellipsis and lexical cohesion (see Kunz and Steiner, 2012, 2013).

In the remainder of this paper we will first discuss several notions of conjunction in the literature in order to arrive at a classification which fits our systemic and textual analyses (Section 2). We will then describe cohesive conjunctions by comparing their (language-independent) features with other types of cohesion, such as reference, substitution, ellipsis and lexical cohesion (Section 3). Section 4 will focus on the systemic comparison of conjunctive strategies of cohesion in English and German. Section 5 will describe the methods applied for our textual comparison of English and German and will present some first findings from our corpus-linguistic analysis. This will finally lead us to conclusions and some suggestions for further work.

2. Cohesive conjunctions – conceptual clarification

As already noted above, there are numerous strands of research dealing with the investigation of conjunction. They not only differ in the languages and registers addressed and the methodologies applied, but first and foremost, in the linguistic levels focused on. With respect to the latter, we can identify three main orientations: 1) Rhetorical, 2) Grammatical and 3) Text-linguistic orientations:

Rhetorical studies consider pragmatic aspects of meaning relations and are concerned with the coherence between textual parts as a communicative function depending on the context of situation. Many of these works start from logic and deal with questions of truth value and validity. Their primary aim is to differentiate types of logico-semantic relations and to describe the relations established between adjacent stretches of text. What distinguishes these works from grammatical and textual studies is that they also consider logico-semantic relations which are not explicitly expressed by linguistic items and which have to be inferred by text recipients on the basis of their conceptual and procedural knowledge of the world. Works in the frame of Rhetorical Structure Theory can be grouped under this orientation (e.g. Mann and Thompson 1987, but also Grimes 1975) as well as Relevance Theoretic works (e.g. Carston 1993, Sperber and Wilson 1986).

Studies with a focus on *grammatical aspects* of conjunctions are interested in the structural and functional properties of explicit linguistic devices that trigger logico-semantic relations. They mainly focus on the devices that combine linguistic elements at different grammatical ranks (conjunctive devices coordinating individual parts of speech, phrases and main clauses subordinating devices). Some of these studies also integrate adverbial constructions as means for linking structurally complex textual parts. Pasch et al. (2003) provide a detailed classification of German conjunctive devices according to their syntactic restrictions. The categories are classified in terms of syntactic functions, whereas subtypes are categorized according to the positional peculiarities of German sentence structure (see below). Blühdorn (2008a) generally distinguishes hierarchical and non-hierarchical connections, which are discussed from the perspective of syntax but also semantics and discourse structure. He argues that the type of discourse connection cannot be automatically derived from syntax and semantics, across languages. Grammars of English more or less use similar criteria as German works for categorizing types and subtypes of conjunction (cf. Quirk et al. 1985 for a detailed approach) although these are not subsumed under a common heading. While few works use the same terminology, three main categories seem to be favored by most of them: coordinating devices, subordinating devices and conjunctive adverbials (see below).

Text-linguistic approaches take conceptual aspects as a starting point and examine conjunctive devices as explicit triggers for logico-semantic relations. They are not so much concerned with a fine-grained differentiation of structural aspects of cohesive devices but rather focus on classifying types of semantic relations. Most of these studies exclude intra-sentential relations signaled by prepositions and phrase coordinators, by particular types of particles, and mostly also by subordinators. Halliday and Hasan (1976) stress one important aspect of cohesive devices: the function "of relating to each other linguistic elements that occur in succession but are not related by other, structural means" (1976: 227). An important point here is that the "cohesive power" (1976: 229) does not rest in a conjunctive expression like *afterwards*, but in the underlying semantic relation. Halliday and Hasan (1976) distinguish several logico-semantic types of relations: additive, adversative, causal, and temporal, which are further subclassified semantically. A fifth category includes "continuatives" (1967: 267ff) and comprises devices such as *well*, *of course*, *after all*, which would elsewhere fall under the category of discourse markers. Martin (1992: 178ff) and also others note that conjunctive relations vary in their textual function in that some are mainly 'rhetorical' (or interpersonal) while others are mainly experiential. Halliday and Matthiessen (2004) point to the fact that logico-semantic relations, especially between sentential sequences often remain implicit although the kind of "semantic relationship is clearly felt to be present" (2004: 548). They suggest excluding implicit relations from a textual analysis as including them would lead to "a great deal of indeterminacy, as regards whether a conjunctive relation is present or not and as regards which particular kind of relationship it is." (2004: 549). In German text-linguistic approaches the idea of "conjunctives" is taken up with the notions of "Konnektive" (Linke et al. 2004) and "junctives" (de Beaugrande and Dressler 1981). De Beaugrande and Dressler (1981) list four main subcategories: "conjunction", "disjunction", "contrajunction", and "subordination".[2] These are comparable to Halliday and Hasan's categories as logico-semantic criteria build the basis for differentiation. Both German works stress the potential of conjunctive devices for creating cohesion sentence-internally (i.e. between phrases, in hypotaxis and in parataxis) as well as between larger textual chunks. Thus sentence boundaries do not play a role in their concepts of cohesion.

Closely related to cohesive conjunction is, according to Halliday and Hasan (1976), the cohesive function of *intonation*. Although they do not include this aspect as a type of conjunctive relation, they suggest it to be "considered as expressing forms of conjunctive relation" (1976: 271). De Beaugrande and Dressler (1981) include intonation as a cohesive strategy as well, in the sense that it is a system that supports cohesion in oral communication.

Our work combines grammatical and textual perspectives since differences on one linguistic level entail or, at least, interact with contrasts on other linguistic levels. We analyse the structural and functional properties of cohesive devices and their connects but also examine the logico-semantic relations. Implicit relations, which lack an explicit linguistic trigger, and also features of intonation are excluded from our research. Our systemic and corpus-linguistic analyses deal with three main structural types of cohesive devices:

- *coordinators*: link textual elements in a paratactic construction (e.g. *and, but, neither ... nor, und, aber, weder ...noch*, etc.)

- *subordinators*: link elements in a hypotactic relation (e.g. *because*, *before*, *weil*, *bevor*, etc.)
- *adverbials*: link clause complexes (sentences) or even elements on a higher textual level (e.g. *therefore*, *by contrast*, *deshalb*, *im Gegensatz dazu*, etc.)

Our interest not only lies in the structural type of the conjunctive devices that signal a relation between two textual elements but also in the logico-semantic relation itself. For our corpus-linguistic analysis we distinguish five main types:

- *additive*: relation of addition, for two events that are true/ not true at the same time (conjunctive devices indicating such a relation are e.g. *and, in addition, und, außerdem*)
- *adversative*: relation of contrast/ alternative, for two events which are not true at the same time (*yet, although, by contrast, doch, obwohl, im Gegensatz dazu*)
- *causal*: relation of causality/ dependence between (*because, therefore, that's why, weil, deshalb, aus diesem Grund*)
- *temporal*: temporal relation between events (*after, afterwards, at the same time, nachdem, danach, gleichzeitig*)
- *modal* (interpersonal): relation between events connected by an evaluation of the speaker (*well, sure, klar, sicher*)

This combined perspective permits identifying contrasts between English and German with respect to the systemic options that are available for signaling logico-semantic relations. In addition, it also allows comparing the options which are realized in texts of both languages by corpus-linguistic analyses. Possible textual variations depend on variation in the context of situation (as reflected by register variation and also written vs. spoken language) and in the type of production (original texts vs. translations).

3. Cohesive conjunction and other types of cohesion

What is common to cohesive conjunctions and all other types of cohesion, at least across English and German, is their potential to establish relations of meaning between textual elements and the fact that the interpretation of the conjunctive devices is dependent on the elements they tie up with. Yet, cohesive conjunctions exhibit several linguistic and conceptual properties which are distinct from other cohesive strategies. Note however that there are borderline cases whose frequencies and functions are assumed to be language dependent.

The following cection will discuss the most typical features of cohesive conjunctions relative to other cohesive types, without going too much into detail about exceptions or peculiarities of individual strategies. We will begin by describing the properties of cohesive devices (Sections 3.1, 3.2 and 3.3) and then will turn to structural and conceptual aspects of the textual elements linked by the cohesive devices (3.4 and 3.5). In Section 3.6 we will discuss the conceptual relations established between conjunctive device and textual elements.

3.1 Syntactic function of cohesive devices

In contrast to other types of cohesive devices, conjunctive devices are rather restricted in terms of their syntactical function. As noted above, there are three main categories: coordinators, subordinators and conjunctive adverbials.

Especially in the German literature, the first two types are not regarded as fully-fledged syntactic constituents as they have a zero position with respect to the two syntactic elements they connect (cf. Pasch et al. 2003, Blühdorn 2008a). Coordinators and subordinators can be differentiated by the types of clauses they link: Coordinators indicate a linkage between main clauses (parataxis, see example 1) whereas subordinators connect a subordinate clause to its main clause (hypotaxis, see example 2).

(1) Peter was ill **and** stayed at home.
(2) Peter stayed at home **because** he was ill.

By contrast, conjunctive adverbials are syntactically integrated into one of their connects and take on the syntactic function of adverbial (as in 3).

(3) Peter was ill. **Therefore**, he stayed at home.

Therefore they are usually not included in classifications of grammatical approaches to conjunction. English grammars mostly distinguish coordinators (e.g. Biber et al. 1999) and adverbials on the basis of their positional restrictions (see below). So-called particles (and focusing particles in particular) have a somewhat exceptional status as their precise syntactic function depends on their position or degree of adjacency. They are therefore either regarded as sentence adverbials or are not considered as fully-fledged syntactic constituents when integrated into the focused constituent (see e.g. König 1991). Although their focusing function is mainly sentence-internal in the latter case, we also assume a cohesive effect and, hence, include particles in our analysis by integrating them into the category of conjunctive adverbials.

In contrast to conjunctive devices, cohesive devices of other types of cohesion are modifiers of noun phrases or prepositional phrases or serve as functional heads (reference and substitution) or lexical heads (lexical cohesion). Therefore, they always are (or are a part of) syntactic constituents and realize all kinds of syntactic functions, e.g. subject, object, adverbial, and even parts of predicates in case of substitution, ellipsis and lexical cohesion.

3.2 Internal structure

Cohesive conjunctive devices exhibit a higher degree of variation than other cohesive devices in terms of their internal structure. They may consist of single parts of speech as in (1), (2) and (3) above.

Coordinators are a closed class of items, consisting of reduced simple elements. As shown in (4), two single parts of speech (POS) may also function as so-called discontinuous or correlative coordinators, where one part of the coordinator is placed before the first connect and the other one before the second connect.[3]

(4) **Either** Peter stayed at home **or** he went to work.

Most subordinators consist of single POS, yet there are also complex subordinators, e.g. *as far as* or constructions with *that* (see e.g. Biber et al. 1999: 85) and correlative subordinators, e.g. *if… then, as … as* (see e.g. Biber et al. 1999: 86).

Conjunctive adverbials exhibit more variation. They may form one single adverb as in (3) or include multiple POS and may formally serve as noun phrases or prepositional phrases as in (5).

(5) Peter was ill. **For this reason**, he stayed at home.
(6) Peter was ill. **That's why** he stayed at home.

Note that we also include constructions in clausal form as in (6), in case they tend towards grammaticalization.

3.3 Syntactic position

As mentioned above, cohesive conjunctives are restricted in their syntactic function, which goes along with some positional features. The most typical position of conjunctive devices, both in English and German, is at the clause boundary between their external and their internal connect, i.e. between the first and the second textual element linked by the conjunctive device. This is illustrated in Figure 1 below.

Positional options vary according to the structural types defined above. Coordinators and subordinators usually introduce their internal connect, i.e. they precede it (see examples above). One exception concerns correlative coordinators and subordinators (see above), where the two connects are introduced by one element.

Internal connects introduced by subordinators may occur in the first position of a hypotactic construction, followed by the external connect representing the main clause (6):

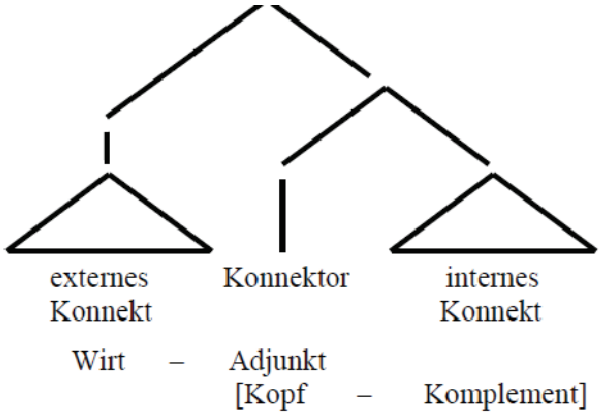


Figure 1. Typical position of conjunctive devices (Blühdorn 2008b: 23)

(6) **Although** Peter was ill, he went to work.

Conjunctive adverbials show more positional variation due to their function as a syntactic constituent. They may either precede the internal connect (see 3 above) or occur at some position inside of it:

(7) Peter **therefore** stayed at home.
(8) Peter stayed at home, **after all**.

Both in English and German one-word constructions show more flexibility in positional options. As mentioned above, the position of particles strongly depends on the position of the focused constituent.

(9) **Even** Peter stayed at home.
(10) Peter **even** stayed at home.

One feature distinguishing conjunctive adverbials from coordinators and subordinators is that they can combine with other adverbials:

(11) There is **however possibly** a different side to the story.
(12) ***And but** there is a different side to the story.

3.4 Number of elements connected

In contrast to the kind of relation triggered by the cohesive devices of reference, substitution, ellipsis or lexical cohesion, conjunctive devices do not trigger a search for one textual element. In fact their relational property is that of “Zweistelligkeit” (Pasch 2003: 3), i.e. they establish a logico-semantic connection between two different textual elements. From a directional point of view they thus are backward- (anaphoric) as well as forward-looking (cataphoric), while other cohesive devices are either anaphoric or cataphoric.

One peculiarity resulting from this property is that conjunctive devices typically realize cohesion rather locally: Unlike devices of (co-) reference and lexical cohesion they are typically not employed in texts to create cohesive chains. However, Martin and Rose (2003: 111) highlight their potential of "sequencing". Most sequences are realized with temporal devices (*first ... then …afterwards ... finally*), but other logical relations may equally be signaled by "chains", such as condition (*if ... as a prerequisite…*) or addition (*and .... or ... and*).

3.5 Complexity of related referents

Cohesive conjunctions are relations between "Sachverhalte" (Pasch et al. 2003) or 'propositions' (e.g. Martin and Rose 2003). Hence the textual elements that are linked by conjunctive devices point to conceptually complex referents like states, processes and events, and not individual entities. This conceptual complexity is reflected linguistically in a relatively high degree of structural complexity of the connects. Conjunctive devices may have a very *wide scope* and link connects that are realized as clause, clause complex or even as textual paragraph (see Martin and Rose 2003: 111, Halliday and Matthiessen 2004: 540). The latter two strategies only apply to coordinators and conjunctive adverbials. In contrast to subordinators, they do not only link propositions but also have an illocutionary force (see e.g. Redder 2003: 497).

(14) ***Although*** Peter is ill, he goes to work.

(15) Peter is ill. He has a high temperature and feels very dizzy. ***Nevertheless*/ *but*** he goes to work.

By contrast, other types of cohesion may also link referents of lower conceptual complexity. Take for instance referential devices such as the neuter pronoun *it*:

(16) We just got back from Paris. **It** took us very long. / **It** was very crowded.

While the first occurrence of *it* refers to the whole event of getting back from Paris, the second instance only refers to the entity *Paris*.

3.6 Type of conceptual relation between related referents

Finally, the meaning relations established by conjunctive devices between conceptual referents differ from those of other cohesive strategies in that they are primarily relational (logical). Other types of cohesive devices are referring expressions in their own right (though often with reduced semantic content) and trigger a search for referring expressions by virtue of their (low) semantic content. They are thus mainly experiential and textual in nature and signal different degrees of similarity between referents.[4] Conjunctive devices do not refer themselves but explicitly signal logico-semantic relations between referring expressions. Again, conjunctive adverbials exhibit some distinctive features. Particularly constructions such as pronominal adverbs carry traces of reference as they contain prepositions linked to a deictic element (*da-*, *hier-*, *wo-* in German). As noted above, they vary however in their 'anaphoric' function. In addition, various multi-word conjunctions include reference items (e.g. *im Gegensatz dazu, that is why*).

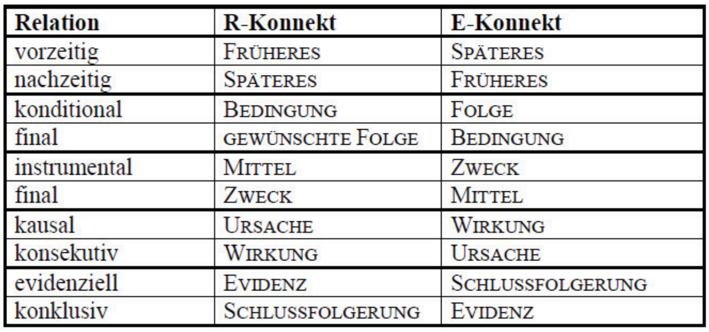
Most studies include the main types of logico-semantic relations such as *additive, spatio-temporal, manner, causal-conditional* etc. Some works provide a more fine-grained model with classifications on different levels of abstraction. For instance, Halliday and Matthiessen (2004: 541ff) define the main logico-semantic parameters of *elaboration*, *extension* and *enhancement* in terms of the basic types of clause-complexing. Each parameter is then classified in two further steps of subcategorization.

Blühdorn (2008b: 26) additionally assigns a thematic role to each of the two connects ("Konnekte"), i.e. the textual elements that are linked by a conjunctive device (see Figure 2).

Variation is possible in terms of linear ordering of the thematic roles. Hence, there are two realizational options for most logico-semantic relations, as exemplified below:

(17) They fought a battle. **Afterwards**, there was a sandstorm.
(18) They fought a battle. **Previously**, it had snowed.

In (17) there is a relation of "Nachzeitigkeit", in which the E-Konnekt (*external* connect) refers to an event *preceding* the event realized with the R-Konnekt (*internal* connect). By contrast, with the relation of "Vorzeitigkeit" in (18) E-Konnekt points to an event that succeeds the one realized via the R-Konnekt.



Relation	**R-Konnekt**	**E-Konnekt**
vorzeitig	Früheres	Späteres
nachzeitig	Späteres	Früheres
konditional	Bedingung	Folge
final	gewünschte Folge	Bedingung
instrumental	Mittel	Zweck
final	Zweck	Mittel
kausal	Ursache	Wirkung
konsekutiv	Wirkung	Ursache
evidenziell	Evidenz	Schlussfolgerung
konklusiv	Schlussfolgerung	Evidenz

Figure 2. Semantic categories and thematic roles (Blühdorn 2008b: 27)

4. Conjunction – systemic differences

We now go on to compare the systemic options that are available in English and German for realizing cohesive strategies of conjunction. If we look into the English and German language systems, the resources for establishing relations of cohesive conjunction appear very similar at first sight. Contrasts between English and German do not reside (so much) in differences with respect to the above-mentioned parameters as such but in the degree of variation within these parameters. We again begin by looking at the properties of the conjunctive devices.

4.1 Variation in syntactic function

Contrasts in terms of syntactic functions between English and German do not so much manifest themselves in terms of availability for the three main structural types coordinators, subordinators and conjunctive adverbials but rather in the potential of particular forms to realize several syntactic functions. We will restrict ourselves to highlighting the main contrasts, focusing on the most distinctive phenomena.

Pronominal adverbs are a very common means for realizing conjunctive adverbial relations in German, but not in English. Due to their internal structure, they may take on various syntactic functions. As a consequence, the relation established by one form may either be mainly cohesive or rather grammatical: Pronominal adverbs may serve as *sentence adverbials* to connect simple sentences or clause complexes, and may thus be cohesive:

(19) Das Betreuungsgeld soll so schnell wie möglich eingeführt werden.
(20) They intend to adopt child care subsidy as soon as possible.
(21) **Dafür** fordert die FDP ein anderes Steuergeschenk.
(22) **For this**, the liberal democrats demand another tax relief.[5]
(23) **Dafür** plädiert die Union schon seit längerer Zeit.
(24) **For this**, the CDU has been advocating for a long time already.

Yet, the same form may trigger a conjunctive or a referential relation of cohesion. For instance, the pronominal adverb in (21) serves as a syntactic adverbial whose meaning is similar to *im Gegenzug* (English: *in return*), while the pronominal adverb in (23) triggers a search instruction for the antecedent, whose scope comprises the whole sentence of (19). Furthermore, *dafür* (English: *for this*) realizes a prepositional object in (23). One translational option would be: *The CSU/ CDU has been advocating this for quite some time.* Hence cases such as (23) are not annotated as cohesive conjunction in our corpus-linguistic analysis but as cases of reference (see Kunz and Steiner 2013).

As a *correlate*, pronominal adverbs mainly have a *grammatical* function. As illustrated in (25), the deictic element refers cataphorically to a proposition which is an argument of the predicate. The pronominal adverb thus introduces a prepositional phrase serving as syntactic object (see e.g. Pasch et al 2003: 491).

(25) Die Union plädiert **dafür**, dass das Betreuungsgeld so schnell wie möglich eingeführt wird.
(26) The CDU advocates **for this**, that child care subsidy should be adopted as soon as possible.

Occurrences of this type are therefore excluded from our corpus-linguistic analysis.

Correlates introducing subordinate clauses are somewhat at the borderline. Here, the pronominal adverb in combination with *dass* functions as subordinating conjunction, as in (27) below.

(27) ... **dadurch, dass** sich ein genetisch neuartiger Typ von der bis dahin gemeinsamen Abstammungslinie abspaltet, bedeutet das auch den Austritt des Neulings aus der bisherigen Erb-Gemeinschaft. [GO_POPSCI]
(28) But when a new species begins to evolve, that is, when a genetically novel type splits off from what till then had been the common line of descent, this means that the new form has left the hereditary community. [ETRANS_POPSCI]

These kinds of correlates are thus assigned to the category of subordinator in our corpus-linguistic analysis.

As already noted above, English and German also differ with respect to the usage of particles. Here contrasts on the one hand concern the inventory of forms, which seems to be richer in German than English, and therefore allow a more fine-grained distinction of meaning relations (see above). In addition, differences between the two languages exist in terms of structural multifunctionality of forms. One cohesive item may indicate a relation between two connects of varying conceptual complexity dependent on its syntactic position. Here, again German seems to provide more options and, in addition, shows more flexibility in combining several particles for different metafunctional purposes (see 4.3 and 4.4 below).[6]

As for English, there are some diachronic changes in the construction of clause complementation that impact on the textual frequency of subordinating conjunctions. One well known peculiarity is the availability of non-finite subordinating clauses with *–ing* participle constructions (see e.g. König and Gast 2012: 74). These clauses may or may not be introduced by a subordinating conjunction. In addition, Mair (2009: 120) points to a rise in the construction of "[i]nfinitival clauses with an explicit notional subject introduced by *for*" (see example [31].

(29) Ministers agreed to take into account the needs of developing countries **while** maintain**ing** the open and nondiscriminatory nature of the multilateral trading system. [EO_ESSAY]

(30) Die Minister vereinbarten, die Bedürfnisse der Entwicklungsländer zu berücksichtigen und gleichzeitig den offenen und nicht diskriminierenden Charakter des multilateralen Handelssystems zu bewahren. [GTRANS_ESSAY]

(31) To restart the unit shortly after powering off, the user must wait for 1 minute **for** the lamp to cool down. [EO_INSTR]

(32) Nach dem Ausschalten sollten Sie mindestens 1 Minute abwarten, bis sich das Gerät abgekühlt hat … [GTRANS_INSTR]

Doherty (1999 and 2004), Fabricius-Hansen (1996,1999) and also House (2011) highlight a tendency towards grammatical metaphor[7] in German, which may translate into a realization of logico-semantic relations at phrase level (e.g. via prepositions). Furthermore, House (2011) and Biber et al. (1999) point to an increase of coordinators in individual registers of written English in order to express relations between events and states in a congruent way. This is interpreted as a reflection of oral style and a general tendency towards colloquiality in Present day English.

Summarizing, we suggest that there are higher proportions of conjunctive adverbials in German than in English texts. English may realize the same or similar meaning relations via other conjunctive devices, or other types of cohesion such as demonstrative reference or more implicitly, by non-finite constructions:

(33) … **Damit** brach eine neue Epoche an.

(34) **This** was the beginning of a new era./ …, **marking** a new era.

As for subordinating and coordinators, we expect higher frequencies in English than German. The same meanings may be expressed in German by conceptually more congruent constructions, e.g. by means of a sentence introduced by a conjunctive adverbial. Alternatively, the logico-semantic relation may be expressed by a preposition in a prepositional phrase, i.e. by a more metaphorical construction.

4.2 Variation in internal structure

The main differences between the two languages under investigation as to the internal form of conjunctive devices seem to result from diverging tendencies of language change. They translate into differences in frequency with respect to forms belonging to the same structural category rather than the availability of the categories as such.

As is commonly known, English exhibits an increasingly isolating morphology with very few inflectional endings left (see e.g. Siemund 2004) while German still provides more inflection and thus has a more fusional character (see e.g. Bittner and Gaeta 2010). These diverging morphological tendencies impact on the structural properties of the conjunctive devices.

English shows a preference for employing multiword constructions in cases where single items would be employed in German. Thus, "extraposed prepositional constructions" and "extraposed absolute linking constructions", as they are called by Bührig and House (2004) realize similar meanings as conjunctive adverbs with derivational endings (e.g. *-lich*) or pronominal adverbs in German:

(35) folglich, damit => as a consequence, as a result
(36) tatsächlich => in fact

German provides a rich repertoire of pronominal adverbs, while in English, fewer forms are available, and most of these are considered dated. Again, English makes use of prepositional phrases (or other complex constructions) to indicate similar meaning relations.

(37) on the contrary => demgegenüber
(38) that is why => deshalb

As pointed out by König (1991: 78) German also has a richer inventory of particles. Fine-grained semantic distinctions realized by individual German particles may correspond to wider meanings expressed by one and the same particle, by syntactic reorganization (e.g. clefting) or via other conjunctive or lexical means in English. According to Miller and Weinert (1999), English usage of particles is mainly restricted to oral communication, while German employs them in written and spoken registers, especially for focusing individual syntactic elements. German also shows a preference for combining several particles:

(39) Peter hat mich angerufen. Klaus hat **aber auch** gratuliert/ **Aber auch** Klaus hat **dann noch** gratuliert.

More prepositional phrases serving as conjunctive devices contain deictic expressions in German than in English.

(40) infolgedessen => as a consequence
(41) im Gegensatz dazu => on the contrary, by contrast

Despite these contrasts opposite tendencies are visible, notably in the register of spoken language. Conjunctive devices with an isolating effect, such as constructions with *that*/ *this* + copula verb + interrogative pronouns (*why/ when/ where*) are contracted in Present Day English when they are spoken. They thus seem to reflect a fusional tendency. By contrast, other constructions seem to suggest a rather isolating tendency in spoken registers of German. For instance, pronominal adverbs quite often occur with a preceding preposition. This however seems to be restricted to combinations of the preposition *von* + pronominal adverb: *von daher/ demher*.

Due to the contrasts discussed above, we generally expect more logico-semantic relations to be signalled by conjunctive devices in German than in English texts. We also suggest that more variation in terms of the forms employed and more multi-word constructions occur in English than German texts.

4.3 Variation in syntactic position

Several factors are known to impact on the degree and type of syntactic variation of cohesive conjunctions and, as a result, on their position in relation to their connects. Here again, we have to draw a general distinction between coordinators, subordinators and conjunctive adverbials. As noted above, coordinators and subordinators are more restricted than adverbials, in both languages under investigation.

Differences are mainly expected with respect to availability of different forms and overall frequency and thus can only be identified on the basis of a corpus-linguistic analysis. For subordinating conjunctions we further assume language-specific preferences for placing subordinate clauses (i.e. the subordinator and internal connect) before or after the main clause (the external connect). We suggest that these preferences are a matter of information packaging (see e.g. Fabricius-Hansen 1996, 1999, Doherty 2004). However language-specific strategies and options of information distribution mainly apply to conjunctive adverbials as they take on the function of a syntactic constituent in the sentence which realizes the internal connect. We suggest that three main factors influence the degree of variation possible in English and German: first, the *internal structure* of the conjunctive device, second, the general syntactic constraints impacting on *constituent order*, and third, the *extent of the scope* of connects.

The *internal structure* of conjunctive devices impacts on their linear syntactic order in that heavy conjunctions generally have fewer positional options than light constituents. Hence, the availability of more light forms (one-word expressions) in German than English should generally coincide with more positional variation. However, *syntactic constraints* partially counteract this flexibility. For instance, more options exist in English for positioning conjunctive adverbials at the left periphery of the clause, somewhere before the finite verb. German only provides one position before the verb. Hence, if a conjunctive adverbial is realized in sentence-initial position, the subject or other main constituents move into the so-called Middlefield, the position between the finite and the non-finite part of the verb complex. English has one particular option in medial position, between operator and main verb, which is only allowed for insertion of single word adverbs (no PPs, etc):

(42) He has **however** told me that …

English additionally permits individual light items such as *as well* and *too* in final position, while only heavy constituents (subordinate clauses) are allowed in German.

By contrast, the German verbal bracket opens up more positional options in the Middlefield. These positions close after the finite are not available in English.

(43) Er relativierte **deshalb/ dagegen/ darüberhinaus** seine Behauptung.[8]
(44) # He relativized **however** his claim.

In addition, there are more options in German for inserting subordinating clauses at some point in the Middlefield. These may be accompanied by elliptical constructions as in (45).

(45) Sie schicken, **wenn** nicht einen Abteilungsleiter, so doch einen kompetenten Vertreter. (Pasch et al. 2003: 362)
(46) # They will send, **if** not a head of department, but instead a competent representative.

German also provides a "Nacherstposition" (the position immediately after the first constituent and before the finite verb) for particular adverbials (see e.g. Becher 2011):

(47) Deutschland **dagegen** hat …
(48) # Germany **by contrast** has …

So called "Postponierer" are only allowed in the position of "Nachfeld" (after the Middlefield or the non-finite verb):

(49) Das ist Stoff auch für Studiengebühren und Stadttheater, **zumal** die Moral immer unmißverständlich rüberkommt. (Pasch et al. 2003: 430)
(50) This also holds for tuition fees and communal theaters, **especially because** the essence of it is …

These diverging syntactic properties are known to entail differences in the distribution of information according to principles of accessibility, relevance and salience, which may also impact on the position of cohesive conjunctions.

One third factor influencing the position of cohesive conjunctions is the *extent of scope***,** i.e. the structural complexity of the internal and external connect, which mostly depends on the conceptual complexity of the referents expressed by the connects. There are various conjunctive devices both in English and German that can be employed for linking connects of varying structural complexity and hence for relating referents of varying conceptual complexity. For instance, some devices serve as a conjunctive adverbial and combine propositions at one position, or rather function as focusing particles, adjacent to the focused constituent, at another position (e.g. *Peter ist* ***auch*** *gekommen vs.* ***auch*** *Peter ist gekommen*).

König (1991) points to the unacceptability of German particles as the only element in preverbal position (e.g. # ***auch*** *ist Peter gekommen*). Büring and Hartmann (2001) observe a stronger adjacency constraint on German than on English particles. Several studies point to a clear contrast in the frequency of particles in English and German. This mainly has to do with the more flexible constituent order in German. It quite easily enhances for instance topicalization in combination with particles for focusing particular constituents in cases where cleft constructions would be favoured in English (see e.g. Ahlemeyer and Kohlhof 1999, Doherty 2004).

Finally, more adverbial connectives are expected in sentence-internal position in German than in English. Becher et al. (2009: 137) postulate that the preverbal position is preferred for propositional parts.

4.4 Conceptual variation

The differences between English and German in terms of conceptual variation involve explicitness or specificity of the cohesive device in terms of the type of logico-semantic relation established as well as the conceptual complexity of the connects. Note that the discussion of fine-grained contrasts is beyond the scope of this artcile and has to be postponed to a later stage of the project. However, some general contrasts are postulated here, partially resulting from the contrasts highlighted in Sections 4.1 to 4.3.

First of all we suggest that the richer inventory of devices in German, especially for conjunctive adverbials, permits a finer differentiation of logico-semantic meaning relations. In addition, positional options in German allow the explicit signaling of more delicate degrees of variation in extent of scopethan English. As an example, König (1991) points to the multitude of possible German translations for the English particle *even*. More explicitness in German is also created by deictic elements, e.g. in pronominal adverbs but also other conjunctive adverbials, as these establish stronger ties to the external connect.

We suggest that degrees of explicitness may also become visible in textual instantiations not only as a consequence of systemic differences but also of preferences which depend on the type of situational context (register). As pointed out by House in a number of works (e.g. 1997), textual contrasts between English and German may be a matter of cultural preferences of verbal communication and may result in a generally higher frequency of conjunctive devices of cohesion in German than in English. Cultural preferences in one language but not the other are also expected with respect to the main (meta-)function (see Halliday and Matthiessen 2004) expressed. We suggest differences in frequency in that German may realize more experiential and textual relations as a reflection of content orientation while English may express more interpersonal relations as a manifestation of orientation towards the addressee (as postulated by House 1997).

In addition we suggest textual differences between English and German with respect to the linear ordering of thematic roles as mentioned in Section 4.3 above which may be due to differences in the internal structure (e.g. one word vs. multi-word construction) and to the syntactic position of the conjunctive device (e.g. sentence-initial vs. middle position). In addition, English and German may differ in the overall distribution of different logico-semantic relations.

5. Conjunction – textual differences

The following section presents differences between English and German with respect to the textual realizations of cohesive conjunction. We first provide information about our corpus resources: the types of texts included, the annotations available and the methods applied to extract distributional information on conjunctive devices and conjunctive relations. After that, we focus on the description and interpretation of the information drawn from the corpus. We do not provide a corpus-based analysis of all systemic contrasts elaborated above, as the study is still ongoing. However, the available results provide insights and tendencies which have to be verified and expanded, in terms of other aspects and different granularity, in the future.

5.1 Corpus resources

We use the GECCo corpus for the analysis of textual differences between English and German in the use of cohesive conjunctions, (cf. Kunz and Lapshinova-Koltunski 2011). It is a multilingual corpus which offers a continuum of different registers from written to spoken discourse and comprises ca. 1.3 Mio tokens. This constellation allows capturing differences in frequency and function of cohesive devices between individual registers across both languages. More precisely, the corpus consists of six parts – four written subcorpora: GO (German Original), EO (English Original), GTRANS (German Translation), ETRANS (English Translation), and two spoken subcorpora: EO-SPOKEN (English Spoken Original) and GO-SPOKEN (German Spoken Original). In this study, we concentrate on the analysis of conjunctive relations occurring in the written subcorpora only, which contain texts from eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAYS), fictional texts (FICTION), corporate communication (SHARE), instruction manuals (INSTR) and websites (WEB).

GECCo is annotated with information on tokens, lemmata, morpho-syntactic features, parts-of-speech (pos), chunks, cohesive devices. It furthermore includes sentence, text and register boundaries and extra-linguistic features (meta-information about experiential domain, goal orientation, social hierarchy, type of interaction derived from register analysis (cf. Biber et al. 2009) or Halliday and Hasan 1989). Moreover, originals and translations are aligned on the level of sentences, as shown in Figure 3. Thus, GECCo provides both comparable (EO vs. GO) and parallel (EO vs. GTRANS and GO vs. ETRANS) subcorpora for our analysis. All annotations were produced with semi-automatic procedures, cf. Lapshinova-Koltunski and Kunz (2011).

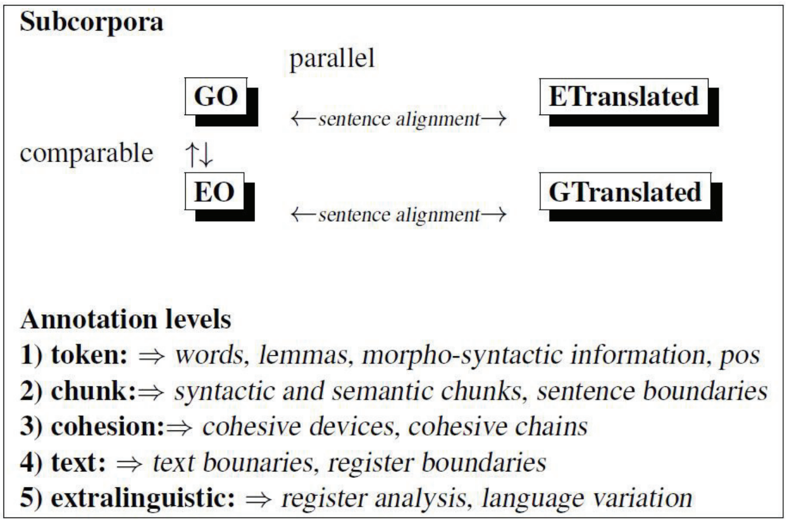
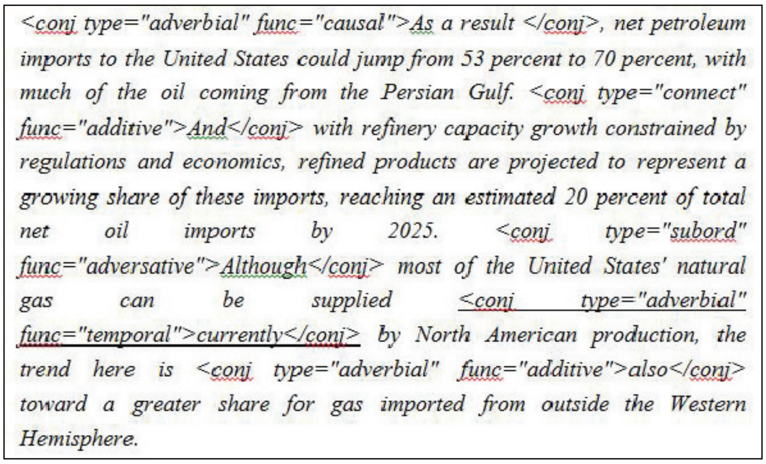

Subcorpora
parallel
GO
←sentence alignment→
ETranslated
comparable
↑↓
EO
←sentence alignment→
GTranslated
Annotation levels
1) token: ⇒ words, lemmas, morpho-syntactic information, pos
2) chunk:⇒ syntactic and semantic chunks, sentence boundaries
3) cohesion:⇒ cohesive devices, cohesive chains
4) text: ⇒ text bounaries, register boundaries
5) extralinguistic: ⇒ register analysis, language variation

Figure 3. Linguistic levels of annotation in GECCo

At the current stage of the project, the annotation of cohesive conjunction comprises information on structural types of the conjunctive device (coordinators, subordinators and adverbials) as well as the main type of logico-semantic relation triggered by the conjunctive device (additive, adversative, causal, temporal and modal), as defined in Section 2 above. The information is encoded semi-automatically with the procedures described in Lapshinova-Koltunski and Kunz (2013) and Lapshinova-Koltunski and Kunz (to appear). These procedures include automatic identification and classification based on the YAC recursive chunker presented in Kermes (2003). The newly annotated information is saved in GECCo as an XML structure *conj* which has two attributes, *type* and *func*, cf. Figure 4. The automatic annotations are manually corrected by humans. This allows us to exclude noise, e.g. cases like *currently* in Figure 4, which were erroneously tagged as conjunctive devices by automatic procedures.



<conj type="adverbial" func="causal">As a result </conj>, net petroleum imports to the United States could jump from 53 percent to 70 percent, with much of the oil coming from the Persian Gulf. <conj type="connect" func="additive">And</conj> with refinery capacity growth constrained by regulations and economics, refined products are projected to represent a growing share of these imports, reaching an estimated 20 percent of total net oil imports by 2025. <conj type="subord" func="adversative">Although</conj> most of the United States' natural gas can be supplied <conj type="adverbial" func="temporal">currently</conj> by North American production, the trend here is <conj type="adverbial" func="additive">also</conj> toward a greater share for gas imported from outside the Western Hemisphere.

Figure 4. Annotation of cohesive conjunctions in GECCo as XML structures

The whole corpus is encoded for use with the Corpus Query Processor (CQP, Evert, 2005), which allows querying annotated structures as well as their sorting and classification. In Table 1, we present queries we applied to compare English and German subcorpora in terms of conjunctive devices.

Table 1. Query for conjunctive devices

	CQP query	**Explanation**
1	<conj>[][]*</conj>	occurrence of all types of conjunctive devices
2	group Last match conj_type	distribution of structural types
3	group Last match conj_func	distribution of semantic types
4	<conj>[]</conj>	occurrence of one-word constructions
5	<conj>[][]+</conj>	occurrence of multi-word constructions
6	group Last match pos	parts-of-speech of conjunctive devices
7	<sentence><conj>[][]*</conj>	occurrence of all devices in sentence initial position
8	group Last match text_register	distribution of devices across registers

CQP enables one to extract information on their overall distribution across subcorpora (query 1), frequencies of their syntactic (query 2) and semantic (query 3) types, internal structure (queries 4 and 5), parts of speech involved (query 6), their position (7) and context of their occurrence (8).

5.2 Corpus-based results and their interpretation

The following section is concerned with the discussion of distribution results extracted from our corpus in accordance with the assumptions discussed above. The figures given below are obtained from the manually corrected corpus.

5.2.1 Overall distribution

We start by examining the general distribution of cohesive conjunctions in our corpus. Figure 5 illustrates the distributions for the two subcorpora: English Originals (EO) and German Originals (GO), which are normalized per 1 million tokens.

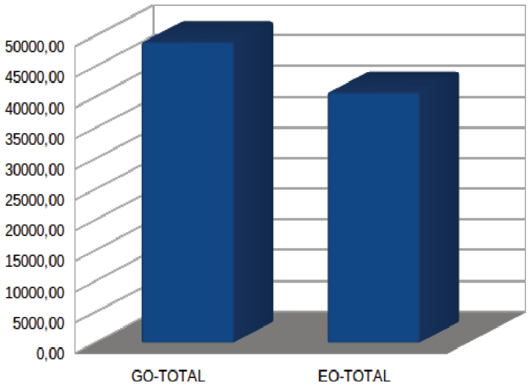


Figure 5. General distribution of cohesive conjunctions

As can be seen from Figure 5, the values of the German corpus exceed those of the English one. These findings corroborate our hypothesis that German favors explicit linguistic means to realize logico-semantic relations in texts. Looking into the parallel corpora reveals that conjunctive devices in the German originals are often left out in their English translations:

(51) Gewerkschaften gibt es in vielen Ländern. Aber nur in den wenigsten ist diese Organisation ein dynamisches Element der Volkswirtschaft. Deshalb irritiert ausländische Beobachter ***auch*** oft der starke Einfluß von Betriebsräten auf unternehmerische Entscheidungen. Doch man kann es ***eben auch*** genau anders herum sehen: [GO]

(52) There are trade unions in many countries. But only in very few of them do these organizations represent a dynamic element in the economy. This is why foreign observers are often confused by the strong influence wielded by works councils on business decisions; yet this argument can be turned around: [ETRANS]

As can be seen above, explicit realizations in the German originals often co-occur with explicit constructions on other linguistic levels (compare *genau anders herum* with *turned around*.

German translations from English show ‘explicitations’:

(53) If we are successful, commercialization of fuel-cell vehicles, hydrogen production, and refueling infrastructure could take place by 2015, … [EO]

(54) Wenn wir erfolgreich sind, könnte **sowohl** eine Vermarktung von Autos mit Brennstoffzellenantrieb **als auch** die Gewinnung von Wasserstoff und eine Infrastruktur zum Auftanken dieser Autos bis 2015 umgesetzt werden, … [GTRANS]

However, we also trace ‘implicitations’ in German translations from English. We observe a slight tendency to leave implicit such meaning relations which have more interpersonal functions as illustrated with examples (55) and (56). By contrast, experiential relations tend to be expressed in German translations.

(55) To some, this display of solidarity may come as a surprise. **After all**, over the past year and a half our respective publics and our media have focused far more attention on … [EO]

(56) Für einige mag diese Bekundung der Solidarität überraschend kommen. In den letzten anderthalb Jahren haben sich die Öffentlichkeit und die Medien in unseren jeweiligen Ländern weit stärker auf das konzentriert, was … [GTRANS]

In what follows, we will show which structural, syntactic and semantic types are mainly responsible for the higher values in German as compared to English.

5.2.2 Syntactic position

Before we go on to explore contrasts in syntactic position between English and German cohesive conjunctions, we will first consider the overall variation in the number of different forms realized in the four subcorpora. By measuring the type-token ratio (percentage of different lexical forms (types) per subcorpus (types/tokens*100 of all cohesive conjunctions), we obtain the findings indicated in Table 2.

Table 2. Type-token ratio of cohesive conjunctions

	EO	ETRANS	ENGLISH TOTAL	GO	GTRANS	GERMAN TOTAL
adverbials	04.27	02.76	01.95	03.48	02.95	04.11
subordinators	01.45	01.58	00.84	02.75	02.40	02.79
coordinators	00.26	00.30	00.14	00.73	00.52	01.74
TOTAL	02.22	01.93	01.18	02.91	02.29	03.37

Table 2 illustrates that fewer differences are identified between the translation subcorpora than between the original subcorpora in that the translations reveal lower type-token ratios than the originals. This is in line with our observations on other linguistic levels (see Steiner and Kunz 2012, 2013) and points to the specific property of Simplification in the translations as compared to their originals, whence the tendency towards using less complex structures (cf. Baker 1996: 176). The type-token ratio in the German original corpus exceeds that of the English corpus by around 0.70. Thus, we find a higher degree of variation in the German original texts as postulated in our hypothesis that German utilizes more types of conjunctions than English. Yet, statistical test are still required to see whether these contrasts are significant. If we take a look at the ten most frequent types (Table 3), we can see that one main cause for the discrepancy between English and German originals is the high number of occurrences of the coordinator *and* in EO (see also below).

Table 3. Frequent conjunctive devices

EO		**GO**	
1892	and	962	und
674	to	654	nur
529	but	567	so
507	when	554	auch
266	also	426	wenn
250	now	402	noch
228	if	391	aber
223	as	338	dann
185	or	290	wieder

We are now in a position to discuss differences between English and German in terms of syntactic position. For this purpose, sentence initial occurrences were contrasted with all other syntactic positions (see Table 4).

Generally speaking we measure the highest proportions for sentence initial conjunctions (distribution in percentage of all conjunctions) in the English original texts (EO), while the lowest proportions are found for the German originals. Furthermore, more differences are traced between the German corpora than between the English ones. In addition, we observe differences between languages in terms of which syntactic types exhibit the highest values occupying sentence initial position: adverbials (24.09%) in English vs. coordinators (17.78%) in German. We now consider each syntactic type in more detail in order to discuss possible influencing factors.

Table 4. Syntactic types in sentence initial position in GECCo

	EO	**ETRANS**	**GO**	**GTRANS**
adverbials	24.09%	21.03%	10.21%	16.24%
subordinators	21.01%	22.67%	14.36%	23.42%
coordinators	18.58%	21.53%	17.78%	15.06%
TOTAL	21.85%	21.58%	12.04%	17.78%

The biggest difference between English and German is found for *conjunctive adverbials* in sentence-initial position. This can be explained by the higher number of pre-verbal options in English, on the one hand, and the availability of several positions in the German Middlefield, on the other.

(57) Ein zentraler Aspekt der amerikanischen Energiepolitik ist **daher** ein Bündel bahnbrechender Technologien, welche die Art der Energiegewinnung und des Energieverbrauchs grundlegend ändern sollen.

(58) **Therefore**, a central aspect of U.S. energy policy is a portfolio of breakthrough technologies that promise to alter fundamentally the way we produce and consume energy.

The figures found for the translations but also the respective instantiations observed suggest an imitation of the source texts (Shining through), which is more pronounced in German than English translations. This is explainable by the fact that English provides fewer options for sentence-internal realizations. For the German translations, sentence-initial position of the cohesive conjunction implies a movement of the main syntactic functions to sentence-internal positions, which may produce a more marked information structure than in the German originals.

As noted above, the sentence initial position of *subordinators* is accompanied by a movement of the internal connect (introduced by the subordinator) before the external connect. Hence, the higher proportions identified for English point to a contrast in the language specific preferences for arranging internal connects before external ones. Comparing the devices employed for initial position, we notice that the most frequent expression in both languages indicates a conditional meaning relation: *if* in English and *wenn* in German. This additionally implies that both languages favor the order of condition before consequence rather than the other way round. One cause for the higher proportion of sentence-initial subordinators in English seems to be related to the realization of non-finite subordinate clauses (as suggested above). For instance, *to* (meaning *in order to*) is amongst the most frequent subordinators in English (76 occurrences), displaying far more occurrences than German *um* (14). However, both languages prefer a linear ordering of external before internal connect.

The figures additionally show an exceptionally high distribution for the German translations. Our extractions suggest that sentence-initial subordinators are, indeed, often employed to translate non-finite clauses (constructions of *to be* and *–ing* participles) which may or may not be introduced by a subordinating conjunction:

(59) **Damit** wir den möglichen Nutzen des Euro realisieren können, müssen wir sicher sein, dass ... [GTRANS]
(60) **To** reap the potential benefits of the euro, we must be sure that ... [EO]

As for *coordinators*, we observe less differing proportions between the original subcorpora than for other conjunctive types. The higher distribution found for English is mainly due to sentence-initial occurrences of *and* as illustrated below. Quite interestingly, more contrasts are identified between the translation corpora than the original ones. Looking at the frequency of each connector reveals that the most frequent expression in sentence-initial position is *and* (220 occurrences) in EO, followed by *but* (207), while it is the other way round in ETRANS with *but* (233) and *and* (181). Most occurrences of *but* correspond to the sentence-initial *aber*, to a lesser extent to *doch,* and also *nur* in the German originals. We also find sentence-internal *aber*, and sentence-initial *but*, and also sentence-internal adverbials indicating a relation of contrast, such as *trotzdem*, *dagegen*, *jedoch*, *dabei*. Omissions or differing structural realizations are very rare.

5.2.3 Internal structure

We now go on to compare the distributions of one- vs. multi-word constructions for conjunctive adverbials. Table 5 shows the distributions in percentages for the four subcorpora and also for originals and translations taken together for each language (ENGLISH TOTAL and GERMAN TOTAL).

Table 5. Proportions of one-word and multi-word conjunctive adverbials (in %)

one word vs. multi	**EO**	**ETRANS**	**ENGLISH TOTAL**	**GO**	**GTRANS**	**GERMAN TOTAL**
adv-one	76.28	76.04	76.14	96.43	96.27	96.36
adv-multi	23.72	23.96	23.86	3.57	3.73	3.64

Here, we observe only minor variations between originals and translations in each language, whereas considerable differences exist between the languages. Extractions from the corpus identify several factors that impinge on the higher distributions for one-word constructions in German: First, they are partially due to the availability of pronominal adverbs. Second, more particles are found in German, especially in the originals. They quite frequently occur in combinations, as can be seen in examples (51) and (54) above. The third factor impacting on the high proportion of one-word adverbs in German is the frequent use of adverbs with derivational suffixes as suggested above. Our analyses show that particular derivational adverbs are employed quite often, serving different textual functions. For instance; *eigentlich* mainly has an interpersonal function, *schließlich* indicates an ideational meaning relation and *zusätzlich* often implies a more textual function. Note that these three adverbials are the most frequent ones and that there is not much variation in derivational adverbs.

As for multi-word expressions, the conjunction *vor allem* (149 occurrences) outnumbers all other multi-word adverbials in German, followed by *zum Beispiel* (57 occurrences) and *darüber hinaus* (21), *unter anderem* (19), *nicht (ein)mal* (19). Hence, the most frequent multi-word adverbials in German are all additive, with a textual or ideational function. There is a considerable frequency gap between the most frequent devices and all others, and except for these five most frequent expressions, none of the other expressions occurs more often than ten times.

The most frequent multi-word expressions in EO are: *that is/ i.e.* (127), *for example* (100), *of course* (59), *at least* (43), *in addition* (42) *That/this is/'s why* (38). Although the two most frequent expressions in English, again, indicate an additive logico-semantic relation and have a more textual or ideational function, we also find frequent adverbials which express other logico-semantic relations and which relate connects on the interpersonal plane. Comparing the English frequencies with the German ones, we also note a more even distribution, with more than 15 types that occur more than ten times. This may imply that a greater systemic potential in German, relative to English, does not necessarily entail more variation in textual realizations.

5.2.4 Syntactic function

Our general observation is a higher frequency of conjunctive adverbials in German than in English because of pronominal adverbs.

Results obtained from the corpus show that the proportion of adverbials among other syntactic types is higher in the German (GO) than in the English originals (EO): ca. 70% vs. ca. 40%. English translations quite often leave out adverbials which are realized in the German originals (e.g. *indes* in 63). In some cases, other cohesive devices are employed instead (e.g. a referential device as in 64), quite often the conjunctive connector *and* is used instead (as in 62).

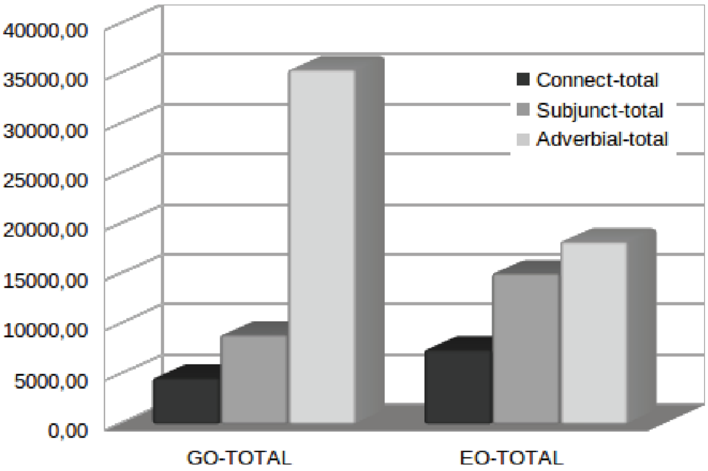


Figure 6. Proportions of syntactic type of conjunctions

(61) Ähnlich geht es **allerdings auch** der Gegenposition, die den Wohlstand für alle dadurcherreichen will, dass … [GO]
(62) **And** the same applies to the opposite view that advocates achieving prosperity for all by … [ETRANS]
(63) Siemens meldete im vergangenen Jahr mehr Patente an als der große Konkurrent General Electric – in dessen Heimatland USA **wohlgemerkt**. Zu Hause hat die deutsche Industrie ihre Investitionen **indes** gestreckt … [GO]
(64) Siemens applied for more patents last year than its famous competitor General Electric – and **this** was in the latter's home country, the United States. At home, German industrial companies have been stretching out. [ETRANS]

The comparison of English originals and German translations shows that occurrences of *and* in EO often correspond with logico-semantically more explicit adverbials in GTRANS.

(65) **And** we also announced: [EO]
(66) **Zudem** haben wir folgende Schritte angekündigt: [GTRANS]
(67) **And finally**, we must encourage energy decisions guided by competitive markets ... [EO]
(68) **Schließlich** müssen wir Entscheidungen im Energiebereich unterstützen, die … [GTRANS]

Furthermore, in those cases, where *and* has a textual rather than an experiential additive function translations mostly do not reflect this, especially when they are combined with adverbials (as in [68]).

5.2.5 Semantic variation

In this last section, we discuss corpus-based results in terms of conceptual variation between English and German. At this stage of our research, we cannot deal with differences in terms of thematic roles of connects. Our extraction results show tendencies in the types of logico-semantic relations established. Figure 7 shows the distributions of different logico-semantic relations in English and German originals.

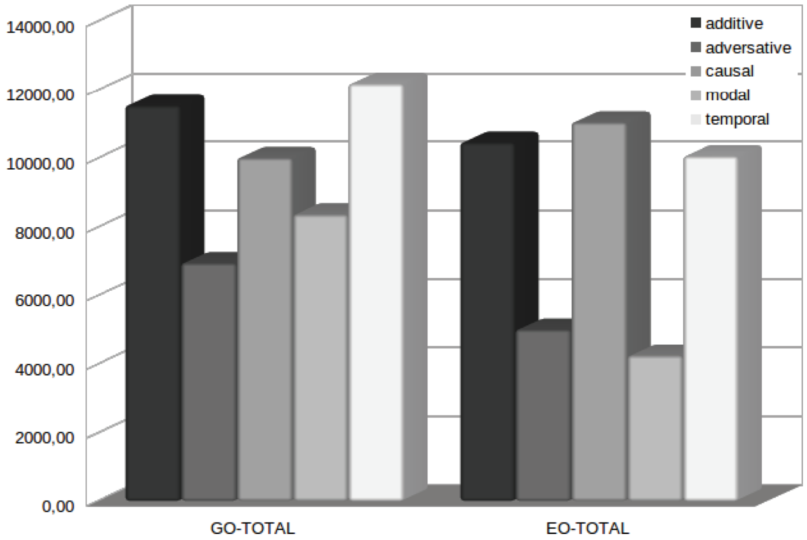


Figure 7. Proportion of logico-semantic conjunctions in GECCo

Generally speaking we observe a preference in English and German for realizing more additive, causal and temporal relations via cohesive conjunctive devices over expressing adversative and modal relations. These quantitative features observed for conjunctive devices have to be treated with some caution. For instance we have already noted above that *and* in English and German *und* are applied with various functions or meanings. In some contexts they have an experiential function, in others they are rather interpersonal.

Additive and *adversative* conjunctions are similar in their function, as they both signal that something new is being added to the discourse (Halliday and Hasan 1976). However most of the additive conjunctions in our English subcorpus signal relations rather vaguely and trigger context-depend relations which can be interpreted on the ideational, textual as well as interpersonal plane. This especially holds for the high distributions of *and*, as already noted above. Higher distributions of adversative relations in the German corpora point to a tendency of argumentation in which different alternatives are evaluated and contradictions explicitated (Christiansen 2011). On the one hand, this reflects general differences between English and German in terms of experiential orientation. However English translations from German originals show that adversative adverbials, particularly with a focusing function, are left implicit. And especially in cases where one adversative item supports the contrastive effect already established by another one, there often is a shift in meaning combined with a syntactic reorganization:

(69) Überfällig **aber** ist **dennoch** diese Debatte über den Sozialstaat in Zeiten des Weniger, in denen Wachstumshoffnungen nicht mehr alles überstrahlen. [GO]

(70) **Nevertheless**, in these lean times when hopes of growth no longer eclipse everything else, this debate is doubtless overdue. [ETRANS]

Interestingly, relations established by adversative conjunctions in German originals are sometimes interpreted as having a purely additive meaning in the translations:

(71) **Dagegen** ist das Gewicht der Bauwirtschaft mit 15 Prozent gegenüber Westdeutschland (4 Prozent) noch entschieden zu hoch. [GO]

(72) **Moreover**, the significance of the construction industry (5 %) remains much too high when compared with western Germany (4 %). [ETRANS]

The contrasts in logico-semantic relations shown above point to another conceptual shift, which concerns the type of thematic role realized by external and internal connect. Although differences between English and German cannot be examined statistically at this stage of our study, we observe some differences when comparing individual extractions:

(73) And **although** the stones seemed simple enough in the midst of the moor, […], he found it was no such thing, […] [EO]

(74) Die Steine hatten mitten im Moor nichts Besonderes an sich gehabt, **doch** nun merkte er, […] [GTRANS]

In this example from our corpus, we find a re-arrangement of conjunctive devices in the German translation (74) although the linear order of the propositions remains. As a consequence, the internal connect of the English original (*the stoned seemed* …) which refers to the "contrasting" proposition is turned into the external connect in the translation (*Die Steine hatten* …), which hence occurrs as 'contrasted' proposition.

Finally, the high distribution of *modal* relations signaled by conjunctions in the German texts primarily results from a frequent use of adverbials such as *eigentlich* and *vielleicht*, which often have a focusing function. English translations often either leave these meanings implicit, as illustrated in (76) or even strengthen them (e.g. by *really* in 78):

(75) Ist **eigentlich** bekannt, dass Deutschland von dieser Vorstellung gar nicht so weit entfernt ist? [GO]

(76) Has everyone forgotten that the German system is not far removed from this concept? [ETRANS]

(77) […] sagt Karl Max Einhäupl, der als Vorsitzender des Wissenschaftsrats den **vielleicht** besten Überblick über die deutsche Forschungslandschaft hat. [GO]

(78) "[…] says Karl Max Einhäupl. And, as chairman of the Science Council, he **really** ought to have the best overview of the German research landscape at the present time. [ETRANS]

Note that the higher distributions of conjunctive modals in German relative to English seem to contradict our hypothesis above, where we have suggested that the tendency towards realizing more interpersonal meaning relations should manifest itself in a higher distribution of modal conjunctives in English than German. However, our English translations point to a tendency for expressing such meanings in predicates rather than adverbials (*forgotten in* 76 and *ought to* in 78). Statistical evidence for this observation has yet to be obtained in translations and originals by a semantic analysis of verbs.

6. Conclusions and future work

In the present study, we combined systemic and textual research and analysed contrasts in strategies of cohesive conjunction in English and German. Our analyses show that although the resources for establishing cohesive conjunction in both languages are similar at first sight, there exist systemic contrasts in the inventory of conjunctive devices, their structural and syntactic properties, the semantic relations they trigger, as well as the semantic and grammatical features of their connects.

These observations are backed up by our quantitative and qualitative corpus-based analysis. Moreover, the extractions from our corpus hint at further variation in form and function, as well as in the semantic relations which will have to undergo in-depth analyses in the future. We could observe differences in occurrences not only between the two languages under analysis but also between text types available in our corpus: originals and translations. This variation demonstrates phenomena of interference which occur in the translation process.

However, we need to undertake further quantitative analyses, e.g. calculation of p-value to evaluate and compare the degree of this variation. Further qualitative analyses are required, particularly to trace more fine-grained contrasts in the conceptual properties of the conjunctive relations and to identify their precise logico-semantic nature. For this purpose, we will include an analysis of variation across registers contained in our corpus. Here we expect even more differences, especially between written and spoken registers, and those which appear to be somewhere in between. As noted above, the study is part of a larger research project in which English-German contrasts are examined for multiple other types of cohesion. We therefore intend to relate the findings obtained for conjunction to those found for other types of cohesion, such as reference and substitution (see Kunz and Steiner forthcoming 2012, 2013) with multivariate analyses.

Notes

[1] The project GECCo: German-English Contrasts in Cohesion is supported by a grant from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). We thank our colleagues in the GECCo team, Marilisa Amoia, Katrin Menzel and Erich Steiner, for their assistance and Erich Steiner also for useful comments. Furthermore, we thank Hannah Kermes for providing the necessary perl script for adaptation.

[2] The term "subordination" is somewhat misleading. It concerns semantic and not structural aspects of subordination.

[3] We include correlative coordinators in our corpuslinguistic analysis of connects although *neither*, *nor* and *both* are sometimes analysed as particles in this construction (see e.g. Hendriks 2004).

[4] Reference: identity between individual referents; substitution and ellipsis: usually identity between types of referents, lexical cohesion: similarity between types of referents, see Steiner and Kunz (2012).

[5] Literal translation.

[6] Recent research (e.g. Büring and Hartmann 2001) however shows that German is less flexible than English with respect to the position of focusing particles as well as the type of syntactic constituent adjoined.

[7] I.e. a tendency to realize referents of high conceptual complexity less congruently, with a relatively low degree of structural complexity i.e. *for the solution of the problem* instead of *in order to solve the problem.*

[8] Note however that personal and demonstrative pronouns have to precede conjunctive adverbials in the Middlefield, e.g. *Er relativierte sie/die deshalb/ dagegen/ darüberhinaus*.

References

Ahlemeyer, B. and I. Kohlhof (1999), 'Bridging the cleft: an analysis of the translation of English *it*-clefts into German', *Languages in Contrast*, 2: 1-25.

Baker, M. (1996), ‘Corpus-based translation studies: the challenges that lie ahead’, in: H. Somers (ed.) *Terminology, LSP and Translation*. Amsterdam: Benjamins. 175-186.

Becher, V., J. House and S. Kranich (2009) ‘Convergence and divergence of communicative norms through language contact in translation’, in: K. Braunmüller and J. House (eds), *Convergence and Divergence in Language Contact Situations.* Amsterdam: Benjamins. 125-152.

Becher, V. (2011), When and why do translators add connectives: a corpus-based study, *Target,* 23: 26-47.

Biber, D. and S. Conrad (2009), *Register, Genre and Style*. Cambridge: CUP.

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

Bittner D. and L. Gaeta (eds) (2010), *Kodierungstechniken im Wandel. Das Zusammenspiel von Analystik und Synthese im Gegenwartsdeutschen.* Berlin: De Gruyter.

Blühdorn, H. (2008a), ‘Subordination and coordination in syntax, semantics and discourse: Evidence from the study of connectives’, in: C. Fabricius-Hansen and W. Ramm (eds) *‘Subordination’ versus ‘Coordination’ in Science and Text*. Amsterdam: Benjamins: 49-58.

Blühdorn, H. (2008b), *Syntax und Sematik der Konnektoren: Ein Überblick*. Mannheim: Institut für Deutsche Sprache, Manuskript.

Bührig, K. and J. House (2004), ‘Connectivity in translation: transitions from orality to literacy’, in: J. House and J. Rehbein (eds) *Multilingual Communication*. Amsterdam: Benjamins. 87-114.

Büring, D. and K. Hartmann (2001), ‘V3 or not V3 – An investigation of German focus particles’, *Natural Language and Linguistic Theory,* 19: 229-281.

Carston, R. (1993), ‘Conjunction, explanation and relevance’, *Lingua*, 90: 151-165.

De Beaugrande, R. and W. Dressler (1981), *Introduction to Text Linguistics.* London: Longman.

Dipper, D. and M. Stede (2006), ‘Disambiguating potential connectives’, Proceedings of KONVENS-06, Konstanz, Germany.

Doherty, M. (1999), ‘Clefts in translations between English and German’, *Target*, 11: 289-315.

Doherty, M. (2004), ‘Strategy of incremental parsimony’, *SPRIKreports,* 25. Available online at http://www.hf.uio.no/ilos/forskning/prosjekter/sprik/pdf/md/MDohertyReport25.pdf.

Evert, S. (2005), *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.

Fabricius-Hansen, C. (1996), ‘Informational density: a problem for translation theory’, *Linguistics* 34: 521-565.

Fabricius-Hansen, C. (1999), ‘Information packaging and translation: aspects of translational sentence splitting (German – English/ Norwegian)’, *Studia Grammatica* 47: 175-214.

Halliday, M.A.K. and R. Hasan (1976), *Cohesion in English.* London: Longman.

Halliday, M.A.K. and R. Hasan (1989*)*, *Language, Context, and Text: Aspects of Language in a Social Semiotic Perspective.* Oxford: OUP.

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

Hendriks, P. (2004), '*Either*, *both* and *neither* in coordinate structures', in: A. ter Meulen and W. Abraham (eds) *The Composition of Meaning: From Lexeme to Discourse.* Amsterdam: Benjamins. 115-138.

House, J. (1997), *Translation Quality Assessment*. Tübingen: Narr.

House, J. (2011), 'Linking constructions in English and German translated and original texts', in: S. Kranich, V. Becher, S. Hoeder and J. House (eds) *Multilingual Discourse Production.* Amsterdam: Benjamins. 163-182.

König E. (1991), *The Meaning of Focus Particles: A Comparative Perspective*. London: Routledge.

Kranich, S., V. Becher and S. Höder (2011), 'A tentative typology of translation-induced language change', in: S. Kranich, V. Becher, S. Hoeder and J. House (eds) *Multilingual Discourse Production.* Amsterdam: Benjamins. 9-44.

Kunz, K. and E. Steiner (2012), 'Towards a comparison of cohesive reference in English and German: system and text', in: M. Taboada, S. Doval Suárez and E. González Álvarez (eds) *Contrastive Discourse Analysis: Functional and Corpus Perspectives.* London: Equinox. 219-251.

Kunz, K. and E. Steiner. (2013), 'Cohesive substitution in English and German: a contrastive and corpus-based perspective', in: K. Aijmer and B. Altenberg (eds) *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson.* Amsterdam: Benjamins. 243-246.

Lapshinova-Koltunski, E. and K. Kunz (2011), 'Tools to analyse German-English Contrasts in Cohesion', in: H. Hedeland, T. Schmidt and K. Wörner (eds) *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language technology (GSCL)*. Hamburg: Universität Hamburg. 243-246.

Lapshinova-Koltunski, E. and K. Kunz (2013), 'Conjunctions across languages, registers and modes: semi-automatic extraction and annotation', in: A. Diaz Negrillo and F. J. Díaz Pérez (eds) *Specialisation and Variation in Language Corpora.* Bern : Peter Lang. 77-104.

Lapshinova-Koltunski, E. and K. Kunz (2014), 'Annotating cohesion for multilingual analysis'. *Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Reykjavik, May 26, 2014.

Linke, A., M. Nussbaumer and P.R. Portmann (2004), *Studienbuch Linguistik.* 5th edition. Tübingen: Niemeyer.

Lüngen, H., C. Puskas, M. Bärenfänger, M. Hilbert and H. Lobin (2006), 'Discourse segmentation of German written text', in: T. Salakoski, F. Ginter, S. Pyysalo and T. Phikkala (eds) *Proceedings of the 5th International Conference on Natural Language Processing* (FinTAL 2006). Berlin: Springer. 245-256.

Mair, C. (2009), *Twentieth-Century English. History, Variation and Standardization.* Cambridge: CUP.

Mann, W.C. and S. Thompson (1988), 'Rhetorical Structure Theory: a theory of text organization', *Text*, 8: 243-281.

Marcu, D. (2000), *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge/MA: MIT Press.

Martin, J.R. (1992), *English Text.* Amsterdam: Benjamins.

Martin, J.R. and D. Rose (2003), *Working with Discourse: Meaning beyond the Clause.* London: Continuum.

Miller, J. and R. Weinert (2009), *Spontaneous Spoken Language: Syntax and Discourse.* Oxford: OUP.

Pasch, R., U. Braußе, E. Breindl and U.H. Waßner (2003), *Handbuch der deutschen Konnektoren: Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Berlin: Walter de Gruyter.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985), *A Comprehensive Grammar of the English Language.* Harlow: Longman.

Redder A. (2003), 'C 12 Konjunktor', in: L. Hoffmann (ed.). *Handbuch der deutschen Wortarten.* Berlin: de Gruyter: 483-524

Siemund, P. (2004), 'Analytische und synthetische Tendenzen in der Entwicklung des Englischen', in: U. Hinrichs (ed.) *Die europäischen Sprachen auf dem Wege zum analytischen Sprachtyp.* Wiesbaden: Harassowitz. 169-196.

Sperber, D. and D. Wilson (1986), *Relevance: Communication and Cognition.* Oxford: Blackwell.

Stede, M. (2002).'DiMLex: A lexical approach to discourse markers. *Exploring the Lexicon - Theory and Computation*', Alessandria: Edizioni dell'Orso.

Stede, M. (2008a), 'Local coherence analysis in a multi-level approach to automatic text analysis', *LDV-Forum* (now: JLCL), 2: 1-18

Stede, M. (2008b), 'Connective-based local coherence analysis: A lexicon for recognizing causal relationships', in: J. Bos and R. Delmonte (eds) *Semantics in Text Processing – STEP 2008*. Research in Computational Semantics Series, London: College Publications. 221-237.

Part 3. Second language acquisition

“Anyway, the point I’m making is”: lexicogrammatical relevance marking in lectures

Katrien L. B. Deroey

Ghent University

Abstract

Drawing on the British Academic Spoken English *(BASE)*[1] *corpus, this paper presents an overview of how lecturers mark important and less important discourse using verbal cues. Such relevance markers (e.g.* the point is, remember, that is important, essentially*) and markers of lesser relevance (e.g.* anyway, a little bit, not go into, not write down*) combine discourse organisation with evaluation and can help students discern the relative importance of points, thus aiding comprehension, note-taking and retention. However, until the research reported here was undertaken, little was known about this metadiscursive feature of lecture discourse, and markers found in the existing literature and EAP materials were rather few and typically not based on corpus linguistic evidence. Combining corpus-based and corpus-driven methods, the research started from a close reading of 40 lectures to identify potential markers. These were next retrieved from the whole corpus and in the case of relevance markers supplemented with other approaches yielding further markers. Relevance markers were mainly classified into lexicogrammatical verb, noun, adjective and adverb patterns, while the markers of lesser relevance were classified pragmatically as indications of message status, topic treatment, lecturer knowledge, assessment, and attention and note-taking directives. This account of relevance marking is valuable for EAP practitioners, will interest lecturer trainers, and provides input for experimental research on lecture listening, note-taking and lecture effectiveness. Furthermore, the paper offers insights into the use of discourse markers such as “the thing is”, “anyway”, “I don’t know” and “et cetera” and illuminates the understudied linguistic phenomenon of relevance marking. Finally, it illustrates the importance of corpus linguistic research and touches upon some difficulties pertaining to assigning discourse functions based on an examination of transcripts only.*

1. Introduction

This paper uses the *British Academic Spoken English* (BASE) corpus to give an account of how lecturers indicate the relative importance of lecture points using lexicogrammatical markers. The ability to distinguish between more and less important information is a key factor in successful lecture comprehension and delivery (Tyler 1992, Lynch 1994, Mulligan and Kirkpatrick 2000, Kiewra 2002, Williams and Eggert 2002). As a form of discourse organisation, relevance marking may facilitate students’ comprehension, note-taking and retention of the complex lecture message (DeCarrico and Nattinger 1988, Williams 1992, Jung 2003, Björkmann 2011), helping them discern the big picture by providing a “guiding pathway” in a morass of information (Revell and Wainwright 2009: 216).

It stands to reason that recognising instances of relevance marking can be especially helpful to the growing numbers of non-native speakers of English attending lectures in English, since it allows them to focus their comprehension and note-taking efforts on the main content. Moreover, even native speaker students are reported to have significant difficulties in identifying key lecture points (see Titsworth and Kiewra 2004 for an overview of studies). Nevertheless, until the research reported here little was known about how lecturers signalled (lesser) relevance lexicogrammatically. The few markers included in English for Academic Purposes (EAP) and educational studies and EAP textbooks typically seem to be based on personal intuitions or experience, and only a few corpus linguistic studies shed some light on this metadiscursive feature of lecture discourse (Swales and Burke 2003, Crawford Camiciottoli 2004, 2007).

This paper provides an overview of relevance marking by bringing together findings from my research on relevance markers (Deroey and Taverniers 2012a) and markers of lesser relevance (Deroey and Taverniers 2012b). These findings have a practical use in that they can inform EAP teaching and materials design, lecturer training, and experimental educational research. They further contribute to our understanding of the use of much-discussed discourse markers such as "the point is", "anyway", and "I don't know" in lecture discourse and increase our insight into relevance marking in discourse. The need for corpus linguistic analyses is also illustrated and the limitations of investigations solely based on transcripts are discussed.

2. Relevance marking

Relevance marking encompasses relevance markers and markers of lesser relevance. The term "relevance markers" (adopted from Hunston 1994: 198) is defined here as lexicogrammatical devices which overtly mark verbally or visually presented points as being comparatively relevant, important or significant. The derived term "markers of lesser relevance" is used for markers which signal less relevant, important or significant information.

Relevance marking combines evaluation with discourse organisation: by evaluating lecture points as (less) relevant, important or significant, the lecturer establishes a hierarchy of importance of these points. It is instances which combine both these lecture functions (cf. Deroey and Taverniers 2011) that we are interested in here. The focus is thus on the evaluation of "discourse entities" rather than "world entities" (Thetela 1997, as cited in Hunston 2000: 182). Put differently, these are instances of evaluation where the lecturer acts as a 'text constructor' rather than "informer" (Hunston 2000: 183). In this way, "an important point" is considered an instance of discourse evaluation, while "an important principle" is disregarded because a world entity is evaluated.

Although the concepts of relevance, importance and significance are not identical, they are closely related. This relatedness is apparent from existing labels for these evaluative categories. Thompson and Hunston (2000: 24) posit an evaluative parameter of “relevance/importance”, while Bednarek’s (2008: 16) parameter of “importance”, Giannoni’s (2010: 77) evaluative category “relevance” and Crawford Camiciottoli’s (2004) “audience-oriented relevance markers” all comprise evaluations of importance, significance and relevance. Lemke (1998) further subsumes importance and significance in one value, and Swales and Burke’s (2003: 6) “relevance adjectives” denote both relevance and importance. Importantly, marking discourse as less or more relevant, important or significant is likely to be interpreted similarly by students.

3. Corpus and methods

The relevance markers investigated in this study are taken from the 160 lectures (1,252,256 tokens) of the *British Academic Spoken English* (BASE) Corpus. These lectures are mostly delivered by native speakers of English and distributed across four broad disciplinary groups: Arts and Humanities (ah), Social Studies (ss), Physical Sciences (ps) and Life and Medical Sciences (ls). A subcorpus of 40 lectures (328,161 tokens) was created for the manual search of relevance markers. It is composed of 10 lectures from each disciplinary group and includes as many different subdisciplines as possible. Study level, interactivity and audience size were systematically varied, and all lectures are by different lecturers.

The following description summarises the more detailed methodological accounts in Deroey and Taverniers (2012a) and Deroey and Taverniers (2012b). Initially, the author and a co-researcher independently identified lexicogrammatical indications of relatively (un)important discourse in four lectures of the subcorpus. Subsequent comparison and discussion of the initial findings led to refined inclusion criteria, which were then used by the author to manually search for further cues in the other 36 lectures. Next, instances of the attested potential markers were retrieved from all 160 lectures using the corpus query system Sketch Engine, which contains the BASE lectures.

For the study on relevance markers, additional methods were used to identify further markers. Briefly, the findings from manual inspection were supplemented by items from the BASE word frequency list (≥50 occurrences) and other research on lectures (Swales and Burke 2003, Crawford Camiciottoli 2004). These items were then retrieved from the whole corpus and any relevance markers discovered in the co-text of the concordances generated through the above methods were added, as well as words which were derived from or synonymous with these (cf. Giannoni 2010). The study of markers of lesser relevance has not yet been extended to include these additional approaches. However, I am fairly confident that the present overview of these markers is fairly representative of the larger corpus since the 40 lectures actually yielded the majority of relevance markers. For both studies, representative samples of relevance markers with sufficient context were interrated by fellow linguistics lecturers. Full agreement was reached about what were instances of relevance markers; for markers of lesser relevance a few minor differences persisted, reflecting the greater reliance of such markers on pragmatic interpretation.

It should be noted that I have worked with the transcripts only.[2] The lectures in the BASE corpus have been audio- or videotaped but it would have been prohibitively time-consuming to consult these systematically in trying to decide the function of particular language. As a result, I have not been able to take into account aspects of speech such as pausing, intonation, stress placement, volume and pace, or visual information such as non-verbal communication or slides. Nor was it possible to triangulate findings through, for instance, interviewing discourse participants. The corpus does contain a few interviews with lecturers but these do not illuminate issues pertaining to relevance marking.

4. Results and discussion

The following account of relevance marking presents the main findings from two separate studies on relevance markers (Deroey and Taverniers 2012a) and markers of lesser relevance (Deroey and Taverniers 2012b). Relevance markers were classified into lexicogrammatical patterns depending on their main element (verb, noun, adjective, adverb), while markers of lesser relevance were classified pragmatically as markers of message status, topic treatment, lecturer knowledge, assessment and attention and note-taking directives. The choice for different classification methods is a result of the main aim of these studies as well as the difference in how the two marker types achieve their relevance marking effect. An important aim of this research is to provide an overview of authentic lexicogrammatical devices which can be used by EAP practitioners. Although some patterns could be discerned with markers of lesser relevance too, a pragmatic categorisation seemed to make more sense as these markers depend much more on the context for their interpretation than do relevance markers. This is also the principal reason why I have not quantified markers of lesser relevance.

Some similarities in the use of the two marker types deserve highlighting here. First, they sometimes cluster with other markers of the same (1, 2) or opposite type (2), thus reinforcing their effect or strengthening the contrast between more and less important information.

(1) I ca**n't remember** it's in the textbook but **ignore** that 'cause it's totally **irrelevant** for the actual what I'm going to tell you (lslct029)

(2) Redcrosse has a fight with one of these guys and wins and another guy and loses **et cetera et cetera** but **the point is** what **you have to remember** is that in an allegorical story everybody the h the hero meets is what he has inside him so it's a way of creating a complex psychological figure (ahlct010)

Second, many markers are frequently accompanied by discourse markers. "And" and "but" are common with both marker types, "now" is more frequent with relevance markers, and “so” occurs much more often with markers of lesser relevance. “Okay” is infrequent and usually occurs on its own rather than in clusters with “so” or “now”; “well” is even rarer. The discourse markers help demarcate the boundaries between less and more relevant discourse, reflect the position of many markers at transition points (functioning as “new episode flags”, cf. Swales and Malczewski 2001) and may also have an attention-focusing effect (Brinton 1996, Swales and Malczewski 2001). Additionally, they sometimes help in recognising the relevance marking function of multifunctional items. For example, “anyway” preceded by “but” (3) or “so” (4) is usually a marker of lesser relevance, with the discourse markers signalling the transition from less to more important information or introducing a “summative evaluation” (Swales and Malczewski 2001: 158) while a combination with “and” (5) tends not to lend itself to a relevance marking interpretation.[3]

(3) the doer sorry the deer I’m on this food poi aren’t I the doer I’m obviously anticipating what I’m going to have for tea tonight but **anyway** the doer is invented (ahlct039)

(4) I won’t go into the details but it it was asking various fairly personal questions and the first I heard well the next thing I heard of it was somebody called up from who’d been accosted by one of these students somewhere down in town and had been asked these questions and we got into a little bit of trouble about it and we hadn’t cle I hadn’t cleared it with the committee because it wasn’t I didn’t believe it was going out so **anyway** there are there are that’s a a very obvious problem but you do need to obtain approval there is a market research society code of practice on on asking questions (sslct002)

(5) medical methods looking at people’s brains obviously weren’t very far advanced and **anyway** maybe that’s the wrong place to look I mean Hume’s interested in thoughts and feelings and so on so really the method that we use is of course introspection (ahlct037)

Third, the most frequent markers (e.g. **V**, **MN v-link** [MN = metalinguistic noun; see Section 4.1.2 below], *anyway*, *et cetera*, *a little bit*, *I don’t know*) are typically multifunctional. Because the inclusion and classification of instances of such markers can have a significant impact on results and conclusions, it is especially important to be reasonably certain about their function in a stretch of speech. The lack of paralinguistic and visual clues as well as information that may be gleaned from interviewing discourse participants is particularly problematic here. Like most corpus analysts working with a fairly large corpus of transcripts, I can therefore only hope to provide a fairly representative picture of the use of these devices with regard to the linguistic phenomenon under study. This is an issue that deserves highlighting, as quantification is often an integral part of corpus linguistic research but usually leads to a simplification of classifications and the assignment of instances to these classes by working with probable principal functions, as I also did in the study on relevance markers. Cheshire (2007) presents some strong arguments against this practice of assigning primary functions.

4.1 Relevance markers

The approach and inclusion criteria detailed above yielded 782 relevance markers. These could largely be classified into lexicogrammatical patterns based on their main lexical constituent (a verb, noun, adjective or adverb) and the grammatical structure in which they occur (cf. Hunston and Francis 2000). The level of detail included in these patterns represents an attempt to balance informativeness with transparency and usability for EAP purposes. For this reason I have avoided the proliferation of patterns that would have been necessary had I included elements that were not strictly essential to the relevance marking function, such as postmodifiers. Consequently, "the point is", "the point about that is" and "the point I'm making is" are all classed as instances of the pattern **MN v-link**.[4] Premodification has been included where it is realised by an adjective expressing relevance, so that "the important point is" counts as a separate pattern **adj MN v-link**. Similarly, in verb patterns I have excluded catenatives (e.g. *want to*) or modals (e.g. *should*) so that instances such as "I'm emphasising this", "I want to emphasise this" and "I should emphasise" are simply classified as belonging the one pattern, viz. **1s pers pron V**.[5] A fifth category 'assessment-related expressions' contains lexemes pertaining to assessment which did not form patterns (1%).

Verb patterns predominate (53.6%), followed by noun patterns (36.5%); adjective patterns are infrequent (7%) and adverb patterns are rare (1.9%) (see Appendix 1). Interestingly, the most frequent verb, noun, and adjective are all prototypical lexemes for marking important discourse (for a complete list of lexemes see Deroey and Taverniers 2012a). They are "remember" (229 instances), "point" (121 instances) and "important" (69 instances) (adjective and noun patterns). Note that this count does not include instances of verbs, nouns, and adjectives in subclauses, which were not analysed separately for this study (but see Deroey 2013 for an analysis which does include these). The predominant patterns are the imperative **V** (ca. 33% of all relevance markers) and **MN v-link** (ca. 20%) (see Appendix 1).

4.1.1 Verb patterns

In verb patterns (see Appendix 1) the main constituent or 'focus' (Hunston and Francis 1998: 49) is a mental or communication verb. These occur without a subject in imperatives and *to*-infinitive clauses or take first person (*I*, *we*) and second person pronouns (*you*) referring to the speaker, listeners or both discourse parties jointly.

The verb patterns with mental verbs (350 instances, ca. 84%) are **1p pers pron V**, **2 pers pron V**, **1s pers pron v + 2 pers pron TO-INF** and **V** (see Appendix 1). They are essentially "cognitive directives" (Hyland 2002: 217) and the verbs denote memory processes (256 instances) (e.g. *bear in mind, forget, remember*) (6), direct attention to verbal or visual points (85 instances) (e.g. *pay attention, note, notice*) (7) or refer to knowledge acquisition (9 instances) (e.g. *know, understand*) (8). The most frequent verbs are respectively "remember" (229 instances) and "notice" (35 instances).

(6) we also have to **bear in mind** that land's also rising (lslct040)
(7) **notice** here the eyes follow objects around (pslct035)
(8) you need to **understand** that there is no such thing as proxy consent for adults (lslct019)

Verb patterns with communication verbs (69 instances, ca. 16%) (**1s pers pron V** and **TO-INF**) express the lecturer's verbal action of highlighting discourse. The verbs range in emphatic force from, for instance, the clearly emphatic "emphasise" and "stress" (9), over "point out" (10), to the least emphatic "make a/the point" (cf. Hunston 2002) (11). The main verbs are "point out" (25 instances) and "emphasize" and "stress" (17 instances each).

(9) now as I say I **want to stress** this point that these aren't two separate systems (lslct036)
(10) and that is just to **point out** that alongside the energy methods that there are this trick of estimating where you are and using approximate processes (pslct022)
(11) and I'm going to **make the point** that the yield if you're breeding a crop it's difficult to breed for several things at once (lslct002)

These mental and communication verbs often combine with deontic modals denoting advisability or obligation (e.g. *should, have to, need*) (6, 8), and with catenatives or modals indicating the speaker's volition (e.g. *want*) (9) and intention (e.g. *be going* to) (11) (for more details and examples, see Deroey and Taverniers 2012 a and Deroey 2013).

The predominance of verb pattern markers is largely due to the popularity of the imperative **V** pattern, especially with "remember". This is likely due to a combination of factors. Imperatives are a fairly economical (Swales et al. 1998) and formulaic way of marking relevance, which may make them useful when trying to deliver an informationally dense message efficiently. Moreover, the potential imposition of such cognitive directives (cf. Hyland 2002) is unlikely to stem their use in lectures as the relationship between the discourse parties is one of institutionalised inequality. Finally, like **MN v-link** (see below) and some other frequent markers of lesser relevance (see Section 4.2), **V** is used very frequently by some lecturers.

The vast majority of **V** markers combine "remember" with clausal complementation. However, instances of **V**, particularly with "remember", are potentially multifunctional (cf. Tao 2001) and can be hard to disambiguate using the text only. Paralinguistic clues and information about what students had been taught before (i.e. whether the lecturer was referring to something students knew already) would have been particularly useful here. Looking at the cotext only, however, when the verb occurs utterance-initially followed by *that*-clause complementation (12) a relevance marking reading is more plausible than when it occurs alone as a clause-final "tag" (13) (Tao 2001: 128) or "parenthetical insertion" (ibid.: 127) (14), in which case the lecturer may be checking students' recollection or linking new to previous information. Instances such as (13) and (14) were therefore excluded, although reactivating previous content can naturally make it more prominent in the listener's mind.

(12) **remember** that kinematics is looking just at the motion nothing to do with forces or anything of that kind (pslct018)
(13) this is important at the European level **remember** (sslct025)
(14) now what is being attempted is to generate beta cells the islet cells which make insulin **remember** from the pancreas prepare those in vitro from a pancreas and invue infuse them through the portal vein which takes them into the liver (lslct011)

4.1.2 Noun patterns

In noun patterns (see Appendix 1) the focus is a metalinguistic noun (MN) which encapsulates discourse. Alternative labels for such nouns include "shell nouns" (Hunston and Francis 2000; Schmid 2000), "general nouns" (Halliday and Hasan 1976), "signalling nouns" (Flowerdew 2003), "discourse labels" (Francis 1994), and "anaphoric nouns" (Francis 1986, as cited in Schmid 2000). The main metalinguistic nouns here are "point" (121 instances) and "thing" (63 instances) (see also Swales 2001, Crawford Camiciottoli 2004, 2007).

The prevalent pattern is **MN v-link**. The other noun patterns have adjectival premodification (ca. 32%), mainly by "important" and "key", (**adj MN v-link**, **deic v-link adj MN**, ***there* v-link adj MN)** and/or contain deictics which encapsulate discourse (**deic v-link MN**, **deic v-link adj MN**) or presentational "there" (***there* v-link MN**, ***there* v-link adj MN**). Various nouns and adjectives are reminiscent of conversational speech (e.g. *big*, *go home message*, *bottom line, thing*) (15, 16, 17), while the predominant pattern **MN v-link** is itself a fairly conversational way of marking relevance (Crawford Camiciottoli 2004).

(15) there's a **big** question there about how far you can get if you're an empiricist philosopher of maths (ahlct037)
(16) don't is the **go home message** (lslct033)
(17) but the **bottom line** is that we don't know (lslct034)

MN v-link is the second most frequent relevance marker (ca. 21%) after **V** (ca. 33%). It has a complex pragmatic profile (cf. Schmid 2000, Simpson 2004, Keizer forthcoming) but may essentially be viewed as a focus formula (Tuggy 1996) signalling noteworthy information. In this vein, Schmid (2000: 94) notes that "the thing is" may be used as "an emphatic linguistic gesture intended to stress the relevance of what one says and to invest it with more significance and importance". The main nouns are "point" (75 instances) and "thing" (29). The prevalence of "point" to some extent reflects the point-driven organisation of lecture discourse (cf. Olsen and Huckin 1990) but also results from the relative ease with which its metadiscursive use can be distinguished. A few instances include lexis denoting enumeration, such as "first", "two", "next", "other" and "only". In such cases, relative clause postmodification helps distinguish relevance marking instances (18, 19, 20) from those with only a discourse structuring function, which were not included.

(18) so the **first** point **to note** is water on its own appears not to do anything (pslct005)
(19) the **next** thing **to bear in mind of course** this is this biography of Agricola has its own agenda (ahlct005)
(20) the **only** thing **to point out about those two** again just to reinforce what I said last time whether you do it just the same as this example doesn't matter but do be systematic (pslct022)

Some of the popularity of this "semi-fixed" construction (Miller and Weinert 1998, as cited in Keizer 2012) may derive from the lecturer's reliance on prefabricated chunks. These may be used as "processing short-cuts" (Wray and Perkins 2000: 17) when having to formulate a complex message in real time (Wray and Perkins 2000, Biber et al. 2004). Indeed, **MN v-link** markers are often unmodified, or "idiomatic" (Sinclair et al. 2004, Biber 2006a) (21). The fact that some lecturers use such instances repeatedly (in way that is reminiscent of discourse markers) also contributes to their prevalence.

(21) so the previous slide was a kind of theoretical prediction if you like the rationalisation this is the experimental result I suppose **the thing is** that C-D-S is rather better than we might have thought it was but it's still nothing like as good as T-I-O-two or zinc oxide (pslct006)

Other instances have relative clause (ca. 32%) (22) or prepositional phrase (ca. 7%) (23) postmodification. Notable collocations between the antecedent and the main verb in the relative clause are "thing" with "remember" or "bear in mind" (19) and "point" with "make" (cf. also Swales 2001) (22).

(22) now the **point I'm making** is that the stabilization measures you can do them very quickly (sslct001)
(23) the message **from that slide** really is that you got to get past this point here to get O-H radicals (pslct006)

The marker **adj MN v-link** (64 instances, ca. 8%) is the second most frequent but far less common noun pattern. However, it is probably the noun relevance marker that intuitively comes to mind and is likely more easily recognisable as marking relevance than unmodified instances of **MN v-link**. Interestingly, in this and other noun patterns with premodification by relevance adjectives, relative clause postmodification is common (24, 25).

(24) now the essential questions **that I want to address in talking about this model** are the following (pslct012)
(25) okay the important thing **to note** is that nothing happens to the internal price levels in the Caribbean (sslct009)

4.1.3 Adjective patterns

In adjective patterns (see Appendix 1) the focus is a predicatively used adjective (ADJ) expressing relevance which is linked (v-link) to a deictic (deic) (**deic v-link ADJ**), metalinguistic noun (mn) (**mn v-link ADJ)**, anticipatory "it" (***it*** **v-link ADJ**), or "what" (***what*** **v-link ADJ v-link**). "Important" occurs in over half the instances (ca. 56%). Adjectives are much more frequent as premodifiers in noun patterns (92 instances) than as the focus of adjective patterns (55 instances).

The prevalent adjective pattern is ***it*** **v-link ADJ**, which represents extraposition. The chief adjectives are "important" and "worth". Anticipatory *it* mostly projects non-finite clauses with verbs denoting the speaker's communicative (26) or listeners' cognitive (27) actions.

(26) **it's important** to point out to you that in these early studies there was a lot of controversy (lslct037)
(27) **it's worth** knowing that (pslct025)

This pattern has previously been noted to typically contain evaluative adjectives (e.g. Hunston and Sinclair 2000, Peacock 2011) and can emphasise noteworthy points (Hewings and Hewings 2001). The informational value of the evaluative frame ***it*** **v-link ADJ** is generally low and thus provides the speaker with "an extended opportunity to formulate the message" (Collins 1991: 214) and the hearer with "breathing space which facilitates processing" (Kaltenböck 2005: 146). The relatively long clausal subject renders the highlighted point completely and is extraposed, facilitating processing; the point furthermore receives added salience from its end position. By contrast, **deic v-link ADJ**, the second most common but much less frequent adjective marker, is possibly less listener-friendly or effective since it is left up to the listener to infer the referent of the deictic, i.e. what the important point is.

4.1.4 Adverb patterns

Adverb patterns (see Appendix 1) consist of an adverb phrase in clause-initial position containing an adverb (*essentially, importantly*, *significantly*) expressing a judgement of importance regarding the proposition that follows. The rareness of such markers (1.9%) is perhaps not surprising as stance adverbs in university classroom teaching have been found to generally express epistemic (e.g. *probably*) rather than attitudinal stance (Biber 2006b). It should be noted, however, that it proved difficult to establish the discourse relevance marking function of adverbs using transcripts only. I have taken a conservative approach to including instances, retaining only those where a relevance reading seems likely. Thus (28) was excluded, as the cotext suggests an evaluation of the importance of an event rather than of the following proposition (29). Nevertheless, even with this approach to inclusion, a few instances such as (30) admittedly remain somewhat ambiguous in what is being evaluated.

(28) the most important thing here is that your G-P tutor is going to be giving you feedback every time you talk to a patient your G-P tutor will be watching and listening to what you do and they will give you feedback on how you did it well what worked why it worked and they'll give you suggestions for things to try so it's not knock you feedback it's useful feedback and **even more importantly** at the end of the course they should give you some structured written feedback as well so that you can take that and put it in your portfolio and you've got a record of that (lslct038)

(29) we'll come back to this topic a little later but **essentially** what I'm saying is that you could have a word instruction which is sort of adding two data registers together (pslct007)

(30) and **very very importantly** we've mentioned it before this afternoon we'll be looking at it in a lot more detail your measures we said if you've got a strategy how do you know it's working because you measure it (sslct035)

4.1.5 Assessment-related expressions

A few instances (seven) do not fit any pattern but signal important information by pointing out the likelihood of being assessed on particular content. These assessment-related expressions contain "exam" or "examine".

(31) there may be a question about this in one of your **exams** so it's the type of table you do need to to know (lslct033)

The rarity of references to assessment (1%) seems counterintuitive given their potential for focusing attention and guiding study (cf. McKeachie 1994). A further search with related lexemes, viz. "examination", "test", and "question" did not yield additional markers. However, the finding may perhaps be partly explained by the lecturer's desire to promote understanding and discourage selective study, possibly as a result of lecturer training in the UK (Michaela Mahlberg, personal communication 2012). Alternatively, it could partly be due to corpus composition in that initial and final lectures, which probably contain most assessment information, may be underrepresented in BASE; however, it is impossible to establish how many lectures are initial and final.

4.2 Markers of lesser relevance

The markers of lesser relevance (see Appendix 2) identified in the 40 lectures were classified into five broad types according to how they signal lesser relevance: message status, topic treatment, lecturer knowledge, assessment, and attention and note-taking directives. As will be discussed below, some of these categories overlap to some extent. The markers have been reduced to core lexemes and are illustrated with examples from the whole BASE lecture corpus. An overview of all markers can be found in Deroey and Taverniers (2012b).

Few markers explicitly evaluate discourse as being irrelevant (e.g. *not pertinent*) and only a few have an inherent meaning of lesser relevance (e.g. *incidentally*). Instead, the markers tend to indicate lesser relevance and depend rather heavily on pragmatic interpretation to achieve their effect. They can generally be viewed as "muted signals" (Swales and Burke 2003: 17) which express relevance implicitly or cumulatively (cf. Hunston 2011). Hence, Hunston's observation that "much evaluative meaning is not obviously identifiable, as it appears to depend on immediate context and on reader assumptions about value" (2004: 157) is particularly pertinent here.

It should therefore be stressed that the lexicogrammatical constructions presented here do not necessarily signal lesser relevance. Instead, what I argue is that they have the potential to be thus interpreted in the lecture context. In addition to co-textual features such as co-occurrence with other markers of (lesser) relevance, collocational behaviour, and position in an utterance, their interpretation to an important extent depends on knowledge about the lecture genre. Relevant generic features will be noted in the discussion below, but essentially, lectures can generally be characterised as monologic, semi-planned and time-bound speech events aimed at enabling learning by efficient and effective dissemination of accurate subject information which may be assessed.

4.2.1 Message status

The markers of lesser relevance included in this category either assign a negative value to part of the lecture message in terms of its relevance or demarcate boundaries between more and less relevant discourse (see Appendix 2). Interestingly, instead of evaluating discourse as irrelevant (e.g. *irrelevant, not matter*) (32, 33), lecturers tend to convey a point is not central to the discourse topic by labelling it as subsidiary discourse (34), a digression (35), or as not being the core business of lecturing (36).[6,7]

(32) in this case it's a subtraction so it's minus-two but it's **irrelevant** and the only thing that's important is the multiplicative factor of three (pslct032)

(33) it does**n't matter** what the actual numbers work out at to work out at the important point is that it's a straight line (lslct002)
(34) but exactly how it does it is too much **detail** for us (pslct007)
(35) and as an **aside** if you're interested in in food studies at all then you know that some storage proteins in seeds are powerful allergens (lslct003)
(36) right finally this is today's **joke** anybody guess what that is (lslct004)

The demarcation between the development of the main topic and digressions is indicated by "incidentally", "in passing", "by the way" and "anyway" (Strodt-Lopez 1991, Fraser 1999, Siepmann 2005, Fraser 2009), which "indicate the role of the utterance in the discourse" (Blakemore 1992: 138). "Anyway" (37) is by far the most frequent of these, although "by the way" (38) is also common. The relevance marking function of the latter, however, is much easier to recognise than that of "anyway", which has the most complex pragmatic profile (see Takahara 1998). As a marker of lesser relevance, "anyway" signals the transition from discourse with "secondary importance" to "something earlier which the speaker views as being of primary importance" and which (s)he "directs the addressee's attention back to" (Takahara 1998: 328, Strodt-Lopez 1991, Fraser 2009). By contrast, "by the way", (38) "incidentally" (39), and "in passing" typically introduce less relevant information, although some instances seem to signal an insertion of relevant information which perhaps the lecturer forgot to mention earlier (40).

(37) the issue has come up before now that medical personnel being involved in research have felt pressured to take part in that research because they have felt that if they didn't then their careers may be prejudiced and however you'd say that that's nonsense but **anyway** the point is this is the position they're in so carers and medical personnel are also got to be considered in the in the issue of ethical research (lslct019)
(38) he then makes up two sisters this awakened African who **by the way** in the middle of the poem becomes an Indian (ahlct001)
(39) let me just go back a bit and show you something else I just wanted to emphasize the difference of approach for the moderator **incidentally** those of you who're familiar with sort of participatory research approaches this is this is very much related to that (sslct002)
(40) it's the H-L-A which is being recognized should say **in passing** that H-L-A antigens on the cell surface are never empty they always have a peptide associated with them okay so again the structure is always peptide plus H-L-A (lslct011)

The dismissive function of "anyway" is often reinforced by "but" (37) while its resumptive function is often brought out by "so" (41) (see also Takahara 1998). The earlier discourse may be the topic at hand but also more generally the main exposition. In addition to demarcating asides and subsidiary discourse, "anyway" also concludes interpersonal episodes such as class management (see Deroey and

Taverniers 2011) (for instance, managing lecture delivery, 42) and jokes (43), and sometimes seems a “self-management device” which the lecturer uses to stop him- or herself from pursuing an irrelevant point (42).

(41) it’s still not illegal actually what he did was not illegal just a slight aside I know it’s terrible inappropriate wrong everything like that but it wasn’t it wasn’t and isn’t illegal you don’t you didn’t and don’t have to ask permission from patients to take bits of organs out of them the law’s going to change soon and it will all be illegal and then pathology will cease to exist **anyway** so minimal change (lslct034)

(42) it’s a very strange slide I I I can’t quite see the connection between rejecting H-nought and and all the other points on the slide **anyway** the P-value gives us an idea of I I a probability of interpretation (lslct015)

(43) so I can draw something like this all right oh God isn’t that awful **anyway** [[laughter]] [[laugh]] that is not a rosette or something it’s supposed to be a something you might find on a loo **anyway** okay this does not I s I mean apart from the fact it’s a very poor representation of the thing I’m trying I was aiming at we just understand that this sort of stick man kind of thing represents a man (sslct036)

4.2.2 Topic treatment

Topic treatment markers (see Appendix 2) form discourse organisational statements that topics or aspects thereof are not or briefly covered. Their pragmatic effect chiefly depends on the assumption that important points will be prioritised within the limited time available while less important ones may be omitted or receive limited treatment. The difference between this category and message status markers is not always clear-cut but the latter evaluate discourse as less relevant or signal boundaries between the main exposition and subsidiary discourse or asides.

The main markers conveying that the lecturer does not intend to cover a point (in detail) combine “not” with the verbs “go into/through” and “talk about”. These verbs commonly collocate with “detail(s)” (44, 45) and some instances of “go into” are part of a larger construction with “time” (46). A relevance reading is more likely if the cotext indicates the reason for not covering something is that it is not on the exam or does not fit in with the goals for the current lecture (44) or students’ study level.

(44) we’ll cover this next week so we wo**n’t go into** the details (pslct008)

(45) but I’m **not** going to **talk about** that in detail now (lslct011)

(46) there’s a very interesting case in the twentieth century of precisely that the same can be seen in certain regional areas I have**n’t time** to **go into** that in terms of certain topics like fuel as well (ahlct019)

Other markers in the topic treatment category signal a topic will be covered in few words or little time. These are mostly communication verbs (e.g. *look at, talk about,* *say*) combined with "briefly" and "quickly" (47) or with quantifying expressions (e.g. *a little bit, not much*) (48, 49). Alternatively, "quickly" and more commonly "briefly" occur clause initially as style disjuncts (50). In addition, "brief" and "quick" premodify nouns which often denote subsidiary discourse (e.g. *example, review, summary*) (51). These topic treatment markers are frequently found at transition points, i.e. where topics, details or examples are introduced (47-50) or where text is summarized (51).

(47) then we just have to decide what is the mathematical form okay what is the mathematical form and I'm just going to **quickly** give you some examples okay (pslct013)

(48) the second type of gauge which I'm **not** going to **talk about much** in this these lecture courses the the automatic gauge gauges (pslct028)

(49) well I want to start off first of all by **saying a little bit** about the life of of Max Weber (ahlct027)

(50) **briefly** the cooperation procedure came about for two main reasons (sslct025)

(51) let's have a quick look at a **quick** review of just what happens when going to remind you what just what happens when people get together (sslct019)

It will be noted that many markers in the topic treatment category, especially those denoting limited treatment, may equally and possibly simultaneously function to mitigate the imposition posed on the audience, who are expected to attend to a long monologue on topics of the speaker's choosing. In this regard, it is worth adding that "a little bit" (cf. also Biber et al. 2004, Nesi and Basturkmen 2006), "brief(ly)" and "quick(ly)" often occur with other mitigating devices such as "just" (Mauranen 2004, Lin 2010), "let me" and "let's" (44, 48). One way in which the markers' mitigating function may be distinguished from their relevance marking function is by establishing whether topic treatment is indeed limited. Interestingly, an examination of the extended co-text of a dozen instances with "a little bit", "briefly", or "quickly" revealed that in about half a fairly long exposition on the topic follows. This would suggest that these markers often serve a mitigating function which can be used to soften the imposition on the audience or to hold their attention at critical points in the lecture (e.g. before a break or at the end) by "minimiz[ing] the expectations required from students" (Biber et al. 2004: 395).

In the topic treatment category I have also subsumed the very common "general extenders" (Overstreet 1999) "(and so on) and so forth" and "et cetera" because they suggest the speaker could say more but elects not to. More specifically, they can signal that "the 'more' that might be said has low value or, for any number of reasons, is simply not worth the expenditure of communicative energy" (Overstreet 1999: 134). However, these are multifunctional devices (Overstreet 1999, Cheshire 2007) and like most markers in the other categories are not necessarily intended to convey lesser relevance. Where the omitted information is not (readily) recoverable, a relevance reading seems likely (e.g. a detail is unimportant, exhaustiveness is not needed) (52, 53). Where the information appears relatively easy to recover through shared knowledge and experiences (including visuals) (54, 55), general extenders may also be motivated by an attempt to make efficient use of time and processing resources: the vagueness introduced by the general extenders conveys "there is more and you know what I mean, so I don't have to be more explicit at this point" (Overstreet 1999: 134). In addition, general extenders can help disguise the lecturer's knowledge gaps and memory lapses. It also worth pointing out that, as with some other popular markers of (lesser) relevance, some lecturers use these general extenders very frequently (five times or more per lecture).

(52) it's going to be there and people can look at it and say oh yes you know that's u-huh I see that's interesting I wonder what **et cetera** okay so it's getting a balance between having it attractive enough to look at and enough information to convey to people the basics of the subject area (lslct001)

(53) top police drivers are often put forward as the model of driving so they can perform a range of skills which have been defined by other police drivers as being important to good driving like skid control hazard perception vehicle control **et cetera** (sslct028)

(54) it doesn't take a a rocket scientist to work out that big blood vessels get smaller don't they and branch off in capillaries **et cetera** so the whole classification of the complications of diabetes of small and large vessels is completely arbitrary (lslct032)

(55) so when we plug those numbers in so you just slog through that horrible formula we had last time work out the the appropriate put in X-equals-nought set Y-equals-nought and see what the constants come out **and so on and so forth** then it turns out that that one comes out to be like that (pslct022)

4.2.3 Lecturer knowledge

Lecturer knowledge markers consist of "I" and "not know",[8] "not remember", "forget", or "not understand". They suggest incomplete or imprecise knowledge or recollection (cf. Beach and Metzger 1997) on the part of the lecturer and hence can be viewed as evidentials indicating how much trust the listener should place in the information provided (Deirde Wilson, personal communication 2012). Their potential for signalling lesser relevance hinges on students' likely assumption that lecturers are experts and that they will also have prepared the lecture sufficiently to be able to present key points. Therefore students may interpret these "insufficient knowledge claims" (Beach and Metzger 1997) as meaning the (precise) information is comparatively irrelevant for the lecture.[9]

Various cotextual elements suggest that some instances of these phrases may serve as markers of lesser relevance. They cluster with other markers of (lesser) relevance (1); occur with subsidiary discourse (56); or are followed by a 'dismissive' "but". Furthermore, the marked information is often imprecisely rendered because of vague language (e.g. *or so*, *something like that*) (57). While such vagueness may be variously motivated (cf. Jucker et al. 2003; Mauranen 2004), here it strengthens the impression that the lecturer has precise information but that (s)he thinks it relatively unimportant for the present communicative purpose (cf. Mauranen 2004).

(56) I'll remind you about th the Gillick case those that don't know about Gillick **I can't remember** the exact details of the case but her daughters one of her daughters who was under eighteen but **can't remember** her age she was over sixteen got contraceptive advice from her doctor (lsct019)

(57) they have received them yet within the last I do**n't know** five-hundred years or so (sslct036)

Again, these devices may fulfill other pragmatic functions (cf. Tsui 1991, Beach and Metzger 1997, Pichler 2007) which do not really pertain to relevance marking or do not suggest limited subject knowledge. For example, the prevalent "I don't know" in our corpus often expresses the lecturer's uncertainty about the students' knowledge or experience base (58). Some instances of these devices may also function interpersonally. They may, for instance, serve to elicit a response or stimulate thinking (59); they may act as mitigators of negative assessments (60) and as face-saving disclaimers (Potter 2004 and Wooffitt 2005, as cited in Pichler 2007) which "leav[e] room … to retreat from the original position, if challenged" (Tsui 1991: 621) (61); and they may reduce the asymmetrical relationship between the expert lecturer and student novice (see Caffi 1999, Deroey and Taverniers 2011) by revealing the lecturer's limited knowledge base (61).

(58) I do**n't know** if you've ever seen a film like this (ahlct020)

(59) that's really pretty rare in English I do**n't know** about in other languages (sslct038)

(60) the same can be said of of of of most specialties you know what is the the point of us I do**n't know** that's not for for for me to judge okay (lslct032)

(61) and I do**n't** really **understand** but there is a charge gradient apparently a a across the glomerulus and sorry I have to look this up 'cause I always **forget** yeah most proteins are very negatively charged and they therefore repel each other and also the glomerular basement membrane is is appa apparent I do**n't know** how the hell they do this but negatively charged (lslct034)

4.2.4 Assessment

Markers in this category are rare and suggest that something will not be assessed. Most convey expectations or recommendations for study (e.g. *forget*, *not know, not learn, not remember*) (62, 63, 64) rather than referring directly to exams (64). Given the intricate relationship between lectures, assessment and ultimately academic success, indications that something will not be assessed are likely to be viewed as implying that it is relatively unimportant, at least for that particular course. The markers usually follow a brief presentation of the information and often occur with a justification for presenting the information (62) or for not assessing it and with statements about what does need to be remembered (64).

(62) I mention it partly because you may hear of it but I'd I'd rather like you to **forget** about it and and try to think of it i in this more simple way (lslct033)

(63) when you actually plug in real boundary conditions to the complicated formula life's fairly straightforward but I mean there was **no** sense in which we expect you to **memorize** that formula for exam conditions or anything of that sort it it's just not the sort of thing that's sensible to do you look it up in a textbook if you want it (pslct022)

(64) and do**n't** get tied up in **learning** too many of those numbers just remember the basic principles (lslct026)

4.2.5 Attention and note-taking directives

Markers in this category (see Appendix 2) manage student attention and note-taking (see Deroey and Taverniers 2011) by directing students not to pay (much) attention to a point (e.g. *ignore, never mind, not worry about*) (65, 66) or not to take notes (e.g. *not copy/write down*) (67, 68). Their potential relevance marking function stems from the close relationship between information retention, attention and note-taking. As with topic treatment markers, an interpretation of lesser relevance seems fairly plausible with other markers of (lesser) relevance in the immediate cotext (65) or when the point is identified as subsidiary or beyond the students' study level. However, many note-taking directives are motivated by the fact that the information can be found in, for instance, handouts and websites (68). Here the lecturer appears primarily interested in efficient teaching and note-taking rather than wishing to suggest the information is less important.

(65) and for reasons which again I'm not going to go into these cells or some of these cells are important in things like allergy but **never mind** that's by the way there are of course other cells but these cells from a biological point of view lymphocytes dendritic cells and macrophages you need to know about (lslct036)

(66) so why does G appear in this formula and the answer is very straightforwardly that it's because using it in this form is what we can use with the tables so do**n't worry about** it is it's just an artificial step (pslct022)

(67) next week we're going to look with me at language learning and how it's moved on how we've got from books like this in nineteen-fifty-nine I please do**n't copy** it down An Intermediate Course for Adult Learners of English (sslct003)

(68) the article bits which I had sort of quoted on this on this slide do**n't write** them **down** just look them up on the web or in in the handbook (sslct019)

5. Conclusion

This overview of relevance markers and markers of lesser relevance from the BASE lecture corpus yields several interesting findings. First, it shows that markers of lesser relevance on the whole require much more pragmatic interpretation to arrive at a relevance marking reading than do relevance markers. Furthermore, many markers of lesser relevance can also have interpersonal functions such as mitigation (e.g. *a little bit*), face saving (e.g. *I don't remember*) and signalling shared knowledge (e.g. *et cetera*). Second, in both marker categories, the prevalent markers (e.g. **MN v-link**, **V n/clause**; *anyway*, *I don't know*) have various potential meanings and functions. This poses a challenge for transcript-based corpus investigations: although contextual information can be used to guide the analysis, important paralinguistic and visual clues cannot be systematically examined and findings cannot be triangulated by, for instance, interviewing discourse participants. Third, the value of corpus linguistic research is again demonstrated by the fact that quite a few of the attested markers, including very common ones, may not be among those that would intuitively come to mind. Conversely, some markers which we may expect to be frequent, such as assessment references, are in fact rare.

This leads us to the value of this research, which goes beyond providing a better understanding of relevance marking (in lectures). EAP practitioners can use the variety of authentic markers to create more representative materials, which may be used to help students to also recognize common but possibly less easily recognisable cues. Furthermore, the markers can be incorporated into materials and lessons designed to improve lecture delivery by non-native and native speakers. Educational practitioners may, for instance, be interested in the variety of attested markers and the scarceness of assessment references. Educational researchers could use the cues as input for experimental research assessing their effectiveness in helping students distinguish important from less important information.

Nevertheless, several interesting questions remain unanswered here. One question concerns the extent to which prosodic cues and non-verbal communication support or substitute lexicogrammatical relevance marking. Another question regards the generalisability of the findings to other lecturing cultures, such as American or Australian lectures. Finally, it would be interesting to ascertain if and how relevance marking by non-native speakers differs from that of native English speakers.

Notes

[1] The BASE corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council. The corpus is available from The University of Oxford Text Archive at http://ota.ox.ac.uk/headers/2525.xml.

2 Note that the BASE concordances generated by Sketch Engine only contain indications of voiced pauses, laughter and some other non-linguistic sounds.

3 A reviewer commented that when "(okay) so" precedes "anyway", the latter might be more akin to a marker of relevance rather than lesser relevance since in such contexts "anyway" apparently reinforces the message by introducing "a summative evaluation" (Swales and Malczewski 2001: 158). As I see it, "okay (so)" here indicates a desire to wrap up less relevant talk and "anyway" signals the intention to move on to talk that is more central to the exposition. I would therefore argue that *anyway* is in essence a boundary marker between less and more relevant discourse and that any highlighting effect is only indirectly achieved by establishing this contrast.

4 However, postmodification does vary in its contribution to relevance marking and in Deroey (2013) I have included it in an alternative classification of the relevance markers. For more fine-grained analysis of postmodification in relevance markers see also Deroey and Taverniers (2012a).

5 See Deroey and Taverniers (2012a) for a detailed account of this type of variation.

6 For the present purposes, I have adopted Fraser's non-formal definition of discourse topic as "what the discourse is currently about" (2009: 893).

7 Note, however, that the inclusion of subsidiary discourse and asides can suggest the importance of a discourse topic, as such discourse can facilitate understanding, generate interest and connect new information to students' previous knowledge and experiences (cf. Strodt-Lopez 1991).

8 The BASE corpus does not use transcriptions reflecting reduced phonetic realisations. Consequently, "dunno" is only represented (and searchable) as "don't know".

9 Note that for this study only markers which stood out as possibly marking lesser relevance were included. A case could of course be made for including "I think" (see Kärkkäinen 2012 for a discussion of "I think" as marking stanced digressions) but this would necessitate the inclusion of a great many other hedges, which would have rendered this study unmanageable.

References

Beach, W. A. and T. R. Metzger (1997), 'Claiming insufficient knowledge', *Human Communication Research,* 23: 562-588.

Bednarek, M. (2008), '"An increasingly familiar tragedy": evaluative collocation and conflation', *Functions of Language*, 15: 7-34.

Biber, D. (2006a), *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: Benjamins.

Biber, D. (2006b), 'Stance in spoken and written university registers', *Journal of English for Academic Purpose*s, 5: 97-116.

Biber, D., S. Conrad and V. Cortes (2004), 'If you look at…: lexical bundles in university teaching and textbooks', *Applied Linguistics*, 25: 371-405.

Björkman, B. (2011), 'The pragmatics of English as a lingua franca in the international university: introduction', *Journal of Pragmatics*, 43: 923-925.

Blakemore, D. (1992), *Understanding Utterances: An Introduction to Pragmatics*. Oxford: Blackwell.

Brinton, L. J. (1996), *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Berlin: Walter De Gruyter.

Caffi, C. (1999), 'On mitigation', *Journal of Pragmatics*, 31: 881-909.

Cheshire, J. (2007), 'Discourse variation, grammaticalisation and stuff like that', *Journal of Sociolinguistics*, 11: 155-193.

Collins, P. C. (1991), *Cleft and Pseudo-Cleft Constructions in English*. London: Routledge.

Crawford Camiciottoli, B. (2004), 'Audience-oriented relevance markers in business studies lectures', in: G. Del Lungo Camiciotti and E. Tognini Bonelli (eds) *Academic Discourse: New Insights into Evaluation*. Bern: Peter Lang. 81-98.

Crawford Camiciottoli, B. (2007), *The Language of Business Studies Lectures*. Amsterdam: Benjamins.

DeCarrico, J. and J. R. Nattinger (1988), 'Lexical phrases for the comprehension of academic lectures', *English for Specific Purposes*, 7: 91-102.

Deroey, Katrien L. B. (2013), 'Marking relevance in lectures: interactive and textual orientation', *Applied Linguistics*, doi: 10.1093/applin/amt029.

Deroey, Katrien L. B. and M. Taverniers (2011), 'A corpus-based study of lecture functions', *Moderna Språk*, 105: 1-22.

Deroey, Katrien L. B. and M. Taverniers (2012a), '"Just remember this": lexicogrammatical relevance markers in lectures', *English for Specific Purposes,* 31: 221-233.

Deroey, Katrien L. B. and M. Taverniers (2012b), '"Ignore that 'cause it's totally irrelevant": marking lesser relevance in lectures', *Journal of Pragmatics*, 44: 2085-2099.

Flowerdew, J. (2003), 'Signalling nouns in discourse', *English for Specific Purposes*, 22: 329-346.

Francis, G. (1986), *Anaphoric Nouns*. Birmingham: University of Birmingham.

Francis, G. (1994), 'Labelling discourse: an aspect of nominal-group lexical cohesion', in: M. Coulthard (ed.) *Advances in Written Text Analysis*. London: Routledge. 83-101.

Fraser, B. (1999), 'What are discourse markers?', *Journal of Pragmatics*, 31: 931-952.

Fraser, B. (2009), 'Topic orientation markers', *Journal of Pragmatics*, 41: 892-898.

Giannoni, D. S. (2010), *Mapping Academic Values in the Disciplines: A Corpus-based Approach*. Bern: Peter Lang.

Halliday, M. A. K. and R. Hasan (1976), *Cohesion in English*. London: Longman.

Hewings, A. and M. Hewings (2001). 'Anticipatory "it" in academic writing: an indicator of disciplinary difference and developing disciplinary knowledge', in: M. Hewings (ed.) *Academic Writing in Context: Papers in Honour of Tony Dudley-Evans*. Birmingham: The University of Birmingham. 199-214.

Hunston, S. (1994), 'Evaluation and organization in a sample of written academic discourse', in: M. Coulthard (ed.) *Advances in Written Text Analysis*. London: Routledge. 191-218.

Hunston, S. (2000), 'Evaluation and the planes of discourse: status and value in persuasive texts', in: S. Hunston and G. Thompson (eds) *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP. 176-207.

Hunston, S. (2002), *Corpora in Applied Linguistics*. Cambridge: CUP.

Hunston, S. (2004), 'Counting the uncountable: problems of identifying evaluation in a text and in a corpus', in: A. Partington, J. Morley and L. Haarman. (eds) *Corpora and Discourse*. Bern: Peter Lang. 157-188.

Hunston, S. (2011), *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York: Routledge.

Hunston, S. and G. Francis (1998), 'Verbs observed: a corpus-driven pedagogic grammar', *Applied Linguistics*, 19: 45-72.

Hunston, S. and G. Francis (2000), *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Hunston, S. and J. Sinclair (2000), 'A local grammar of evaluation', in: S. Hunston and G. Thompson (eds) *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP. 74-101.

Hyland, K. (2002), 'Directives: argument and engagement in academic writing', *Applied Linguistics*, 23: 215-239.

Jucker, A. H., S. W. Smith and T. Lüdge (2003), 'Interactive aspects of vagueness in conversation', *Journal of Pragmatics*, 35: 1737-1769.

Jung, E. H. (2003), 'The role of discourse signaling cues in second language listening comprehension', *The Modern Language Journal*, 87: 562-577.

Kaltenböck, G. (2005), 'It-extraposition in English: a functional view', *International Journal of Corpus Linguistics*, 10: 119-161.

Kärkkäinen, E. (2012), 'On digressing with a stance and not seeking a recipient response', *Text & Talk*, 32: 477-502.

Keizer, E. (2012), '*The X is (that)* constructions: An FDG account', paper presented at the Second International Conference on Functional Discourse Grammar, Ghent, Belgium, June 2012.

Keizer, E. (forthcoming), 'The X is (is) construction: An FDG account', in: J. L. Mackenzie and H. Olbertz (eds) *Casebook in Functional Discourse Grammar*. Amsterdam: Benjamins.

Kiewra, K. A. (2002), 'How classroom teachers can help students learn and teach them how to learn', *Theory into Practice*, 41: 71-80.

Lemke, J. L. (1998), 'Resources for attitudinal meaning: evaluative orientations in text semantics', *Functions of Language*, 5: 33-56.

Lin, C.-Y. (2010), '"... that's actually sort of you know trying to get consultants in...": functions and multifunctionality of modifiers in academic lectures', *Journal of Pragmatics*, 42: 1173-1183.

Lynch, T. (1994), 'Training lecturers for international audiences', in: J. Flowerdew (ed.) *Academic Listening: Research Perspectives*. Cambridge: CUP. 269-289.

Mauranen, A. (2004), '"They're a little bit different"…: observations on hedges in academic talk', in: K. Aijmer and A.-B. Stenström (eds) *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: Benjamins. 173-197.

McKeachie, W. J. (1994), *Teaching Tips: Strategies, Research, and Theory for College and University Teachers*. Lexington: Heath and Co.

Mulligan, D. and A. Kirkpatrick (2000), 'How much do they understand? Lectures, students and comprehension', *Higher Education Research and Development*, 19: 311-335.

Nesi, H. and H. Basturkmen (2006), 'Lexical bundles and discourse signalling in academic lectures', *International Journal of Corpus Linguistics*, 11: 283-304.

Olsen, L. A. and T. N. Huckin (1990), 'Point-driven understanding in engineering lecture comprehension', *English for Specific Purposes*, 9: 33-47.

Overstreet, M. (1999), *Whales, Candlelight, and 'stuff like that': General Extenders in English Discourse*. New York: OUP.

Peacock, M. (2011) 'A comparative study of introductory "it" in research articles across eight disciplines', *International Journal of Corpus Linguistics*, 16: 72-100.

Pichler, H. (2007), 'Form-function relations in discourse: the case of "I don't know"', *Newcastle Working Papers in Linguistics*, 13: 174-187.

Potter, J. (2004), 'Discourse analysis as a way of analysing naturally occurring talk', in: D. Silverman (ed.) *Qualitative Research: Theory, Method and Practice*. London: Sage. 200-221.

Revell, A. and E. Wainwright (2009), 'What makes lectures "unmissable"? Insights into teaching excellence and active learning', *Journal of Geography in Higher Education*, 33: 209-223.

Schmid, H. J. (2000), *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: Walter de Gruyter.

Siepmann, D. (2005), *Discourse Markers across Languages: A Contrastive Study of Second-level Discourse Markers in Native and Non-native Text with Implications for General and Pedagogic Lexicography*. New York: Routledge.

Simpson, R. (2004), 'Stylistic features of academic speech: the role of formulaic expressions', in: U. Connor and T. A. Upton (eds) *Discourse in the*

Professions: Perspectives from Corpus Linguistics. Amsterdam: Benjamins. 37-64.

Sinclair, J., S. Jones, R. Daley and R. Krishnamurthy (2004), *English Collocation Studies: The OSTI Report*. London: Continuum.

Sketch Engine. Available online at http://www.sketchengine.co.uk/.

Strodt-Lopez, B. (1991), 'Tying it all in: asides in university lectures', *Applied Linguistics*, 12: 117-140.

Swales, J. M. (2001), 'Metatalk in American academic talk: the cases of point and thing', *Journal of English Linguistics*, 29: 34-54.

Swales, J. M and A. Burke (2003), '"It's really fascinating work": differences in evaluative adjectives across academic registers', in: P. Leistyna and C. F. Meyer (eds) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi. 1-18.

Swales, J. M. and B. Malczewski (2001), 'Discourse management and new-episode flags in MICASE', in: R. Simpson and J. M. Swales (eds), *Corpus Linguistics in North America: Selections from the 1999 Symposium*. Ann Arbor: University of Michigan Press. 145-164.

Swales, J. M., U. K. Ahmad, Y. Chang, D. Chavez, D. F. Dressen and R. Seymour (1998), 'Consider this: the role of imperatives in scholarly writing', *Applied Linguistics*, 19: 97-121.

Takahara, P. O. (1998), 'Pragmatic functions of the English discourse marker "anyway" and its corresponding contrastive Japanese discourse markers', in: A. H. Jucker and Y. Ziv (eds) *Discourse Markers: Descriptions and Theory*. Amsterdam: Benjamins. 327-351.

Tao, H. (2001), 'Discovering the usual with corpora: the case of "remember"', in: R. C. Simpson and J. M. Swales (eds) *Corpus Linguistics in North America: Selections from the 1999 Symposium*. Ann Arbor: University of Michigan Press. 116-144.

Thetela, P. (1997), *Evaluation in Academic Research Articles.* Unpublished PhD thesis, University of Liverpool.

Thompson, G. and S. Hunston (2000), 'Evaluation: an introduction', in: S. Hunston and G. Thompson (eds) *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP. 1-27.

Titsworth, B. S. and K. A. Kiewra (2004), 'Spoken organizational lecture cues and student notetaking as facilitators of student learning', *Contemporary Educational Psychology*, 29: 447-461.

Tsui, A. B. M. (1991), 'The pragmatic functions of "I don't know"', *Text*, 11: 607-622.

Tuggy, D. (1996), 'The thing is that people talk that way. The question is why?', in: E. H. Casad (ed.) *Cognitive Linguistics in the Redwoods: The Expansion of a New Paradigm in Linguistics*. Berlin: Mouton de Gruyter. 713-752.

Tyler, A. (1992), 'Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse', *TESOL Quartely*, 26: 713-729.

Williams, J. (1992), 'Planning, discourse marking, and the comprehensibility of international teaching assistants', *TESOL Quarterly*, 26: 693-711.

Williams, R. L. and A. C. Eggert (2002), 'Notetaking in college classes: student patterns and instructional strategies', *The Journal of General Education*, 51: 173-199.

Wooffitt, R. (2005), *Conversation Analysis and Discourse Analysis: A Comparative and Critical Introduction*. London: Sage.

Wray, A. and M. R. Perkins (2000), 'The functions of formulaic language: an integrated model', *Language & Communication*, 20: 1-28.

Appendix 1

Relevance markers: patterns, examples, raw frequencies and percentages (N=782)[1,2]

Pattern type	Example	N	%
Verb patterns		**419**	**53.6**
1s pers pron V	and I want to **emphasize** this (ahlct034)	70	9
1p pers pron V	we should **bear in mind** the ambiguities of the time (ahlct002)	30	3.8
2 pers pron V	but for the moment if you simply **note** that the new industrial organization approach put particular emphasis on this notion of commitment (pslct012)	41	5.2
1s pers pron v + 2 pers pron TO-INF	so I'd like to **ask** you to **pay attention** to both ways of dealing with the problem (pslct034)	9	1.2
TO-INF	this is just a recap from last week just to **emphasise** the changes that are likely to happen in the next hundred years (lslct040)	6	0.8
V	**remember** that Greece doesn't exist at this time (ahlct002)	263	33.6
Noun patterns		**285**	**36.4**
deic v-link MN	these are definitely **things** that you need to know (ahlct010)	23	2.9
deic v-link adj MN	that's one of the big **questions** that we'll continue to come to to come back to (sslct009)	25	3.2
MN v-link	now the **point** is sound waves don't really interact very much with the atmosphere (pslct027)	162	20.7
adj MN v-link	but the key **thing** to note is that it's much nicer to write down (pslct010)	64	8.2

there v-link MN	if there is one **message** that I want to gev gi get across in this course it is that everything helps everything else (lslct002)	8	1
there v-link adj MN	there are two main **ideas** that you need to keep in mind (ahlct024)	3	0.4
Adjective patterns		**55**	**7.1**
deic v-link ADJ	this is so **important** to remember (lslct005)	14	1.8
mn v-link ADJ	that point about elitism is quite **important** (sslct003)	2	0.3
it v-link ADJ	it's **significant** to remember that at the time nuclear power was seen as one of the major potential motors for development in lesser developed countries (sslct020)	35	4.5
what v-link ADJ v-link	but what is **important** to grasp is that our immune system can actually respond to almost anything any protein (lslct036)	4	0.5
Adverb patterns	but **significantly** he's observed from the outside (ahlct008)	**15**	**1.9**
Assessment-related expressions	there may be a question about this in one of your exams (lslct033)	**8**	**1**
Total		782	100

Notes

[1] As a result of ongoing work with the data, there are a few very minor differences between the frequencies reported here and those originally reported in Deroey and Taverniers (2012a).

[2] To represent the surface patterning of similar concordances, Hunston and Francis' (2000) notation system is used: the pattern focus is in upper case, other elements are in lower case and lexemes are in italics.

Appendix 2

Markers of lesser relevance: types, subtypes and examples.

Type	Examples
Message status	
Assigning a negative value	this will be a very brave physician that will say okay we've got laboratory information that tells us the patients are tolerant therefore we shall stop the drugs so there are ethical issues here again all right so that's a little bit of a sideway **sideline** why would we want to discontinue drugs (lslct011)
Boundary demarcation	there'll be three presentations by students and I think you're let off some essay if you do this but we already have do we already have papers don't know but **anyway** there's three there's three presentations (ahlct036)
Topic treatment	
No coverage	I do**n't** want to **look at** the export side at the moment (sslct009)
Limited coverage	we've got two other resource types that I want to look at sorry three other fairly **briefly** the next is a **quick** look at mineral resources (pslct001)
Lecturer knowledge	all you want is a I do**n't know** histogram of sales by region or something (sslct032)
Assessment	the other thing the O-N-B which is all you need to know do**n't** try and **learn** the long title what it's doing in Rome is it's setting up plans (ahlct004)
Attention and note-taking directives	
Attention directives	but **ignore** that 'cause it's totally irrelevant for the actual what I'm going to tell you (lslct029)
Note-taking directives	I'll give you the first few observations not necessarily all of them I mean I'll write them all down but you do**n't** need to **copy** them all down (pslct039)

Faux amis in speech and writing: a corpus-based study of English false friends in the production of Spanish students

María Luisa Roca-Varela

University of Santiago de Compostela

Abstract

The crosslinguistic phenomenon of faux amis *has been extensively studied in different fields of language research, such as translation (Granger and Swallow 1988, Venuti 2002, Malkiel 2006, Ruiz Mezcua 2008), lexicography (Hill 1982, Cuenca Villarejo 1987, Prado 2001, Postigo Pinazo, 2007), and second language acquisition research (Lengeling 1995, Frutos Martínez 2001, Wagner 2004, Chacón Beltrán 2006).* Faux amis *(Koessler and Derocquigny 1928), also referred to as "false friends" (Zethsen 2004, Chacón Beltrán 2006) or "deceptive cognates" (Lado 1957, Batchelor and Offord 2000) are words which share similar forms in two or more languages but have different meanings and/or uses in each language (e.g. English* carpet *'rug' versus Spanish* carpeta *'folder'; English* fabric *'cloth' versus French* fabrique *'factory'; German* Gift *'poison' versus English* gift *'present'). Despite the wide range of surveys, there is a conspicuous scarcity of studies which apply a corpus-based methodology to the investigation of these words; and none of the existing corpus-based studies explore the presence of English false friends in spoken learner language (Granger 1996, Palacios Martínez and Alonso 2005). The present study aims at filling this void by examining 100 high-frequency English false friends in the spoken and written performance of Spanish learners of English through an analysis of two learner corpora, namely the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI). The data obtained from these corpora allow us to draw conclusions about the learners' active use of these lexical items in speech and writing. A total amount of 1403 sample sentences have been closely examined. My analysis reveals that EFL learners make more errors with false friends in their written than in their spoken production and it also shows that certain English false friends are especially difficult for learners (e.g.* actually, pretend, argument*). Thus, the findings of this study certainly shed some light on students' problems in this lexical area which should be addressed in an EFL context.*[1]

1. Introduction

The English language has a significant number of Latin-based words in its lexical repertoire. These words are a remnant of the strong impact of Latin and French on the history, development and evolution of English. As a consequence of this, there are noticeable lexical similarities between the lexicons of many Romance languages and English (Granger 1996). This lexical likeness may sometimes help students in their learning of English vocabulary; however, this is not always the case. Thus, there are words which look alike in English and other Romance languages, and yet, they do not have the same meaning, as illustrated by the so-called *faux amis* (Koessler and Derocquigny 1928), deceptive cognates (Lado 1957) or false friends (Hill 1982).

The existence of false friends between languages has attracted the attention of many linguists, professional translators, lexicographers and language teachers (Hill 1982, Granger and Swallow 1988, Prado 2001, Venuti 2002, Frutos Martínez 2001, Chacón 2006). However, few of them have conducted corpus-based studies on this crosslinguistic issue with the exception of Granger (1996), and Palacios and Alonso (2005). These three scholars analysed learners' writings in order to identify their difficulties with English false friends. Granger analysed the production of learners in the French component of ICLE and reached the conclusion that one third of the lexical errors found involved the misuse of a false friend (e.g. *The economic objective required a* ***unique*** *currency* meaning "single, common"). On the other hand, Palacios and Alonso explored 25 false friends which differ completely in meaning in English and Spanish (e.g. English *disgust* 'repulsion' versus Spanish *disgusto* 'sadness') and concluded that these lexical items cannot be overlooked in EFL classrooms. Although the relevance of these studies cannot be denied, they show some important gaps which need to be addressed.

The present study aims at filling the lacunae left in previous surveys by exploring 100 high-frequency English words which are false friends with European Spanish. This survey looks into the learners' use of these 100 false friends in both spoken and written production on the basis of the data provided by two learner corpora: ICLE and LINDSEI (the *International Corpus of Learner English* and the *Louvain International Database of Spoken English Interlanguage*, respectively). It pays particular attention to the learners' difficulties with these words with a view to identifying the major problem areas in the learners' productive use of English. After explaining the motivation and the main aims of this study, I will describe the procedure followed, the materials used and the results obtained. Finally, I will formulate the main conclusions and pedagogical implications that can be drawn from the analysis.

2. Aims of study

This corpus-based study examines the use of 100 English false friends by Spanish advanced learners. It aims at revealing how well students actually know and use these lexical items in English and to what extent their misuse may influence the quality of the learners' performance. In this way, this study wants to contribute to identifying learners' needs with regard to these lexical items and to devising teaching strategies that can meet these needs in the English classroom.

The main research aims of this study can be summarised in four general objectives:

- to analyse the occurrences of 100 high-frequency false friends in corpora containing speech and writing by advanced Spanish learners of English;

- to identify the lexical errors in this dataset;
- to compare and quantify the errors in the spoken and written data to assess whether there are significant differences in the amount of errors found in the learners' spoken and written performance;
- to identify those English false friends which are particularly challenging for learners.

3. Data and methodology

To realize the aims specified in Section 2, a number of methodological decisions had to be made. The first step was to compile a list of English lexical items with deceptive cognates in Spanish which are important for language learners (Table 1). The main criteria were the frequency of the items in English and their inclusion in authoritative sources. These criteria were implemented into two conditions that seemed useful and practical from the perspective of EFL teaching and learning. The selected English items

(i) had to be *high-frequency words* listed in renowned frequency word lists (the *Longman Communication* 3000 *Word List* and Kilgarriff's word list);[2]

(ii) had to be included in at least four out of the following five specialised reference works: Hill's *A Dictionary of False Friends*, Cuenca's *Diccionario de érminos equívocos ("falsos amigos") inglés-español-inglés*, Prado's *Diccionario de falsos amigos: inglés-español*, Walsh's *False Friends and Semantic Shifts* and Postigo-Pinazo, *Diccionario de falsos amigos: inglés-español*.

From a total of around 12,000 possible candidates, the one hundred false friends listed in Table 1 were selected. They are high-frequency words (with a ranked frequency between 0 and 6,300) which are worth knowing and useful in different contexts. After the initial selection of these false friends, two learner corpora were used to analyse their use in the interlanguage of Spanish learners: the *International Corpus of Learner English (ICLE)* and *the Louvain International Database of Spoken English Interlanguage (LINDSEI).* These two learner corpora are fully comparable since the compilers applied the same criteria. ICLE contains written data (Granger, Dagneaux and Meunier 2002), while LINDSEI represents spoken language (Gilquin, Cock and Granger 2010). ICLE is a computerised corpus of argumentative essays written by advanced EFL learners from widely different L1 backgrounds. The Spanish national subcomponent contains 200,376 words, consisting of argumentative essays of between 500 and 1000 words written by advanced EFL learners, typically university students in their 3rd or 4th year of English studies at the Universidad Complutense de Madrid. LINDSEI can be said to be the spoken counterpart of ICLE. LINDSEI contains oral data produced by advanced learners of English from eleven different mother tongue backgrounds. The Spanish subcomponent contains 50 interviews, and consists of 118,536 words.

Table 1. List of false friends under analysis

ACCOMMODATE (VB)
ACTUAL (ADJ)
ACTUALLY (ADV)
ADEQUATE (ADJ)
ADVERTISE (VB)
ADVICE (N)
ADVISE (VB)
ANNOUNCE (VB)
APPARENT (ADJ)
APPOINT (VB)
ARGUMENT (N)
ASSIST (VB)
ATTEND (VB)
BALANCE (N)
BANK (N)
BATTERIES (N)
BIZARRE (ADJ)
BLANK (N)
CAMP (N)
CAREER (N)
CARPET (N)
CASUAL (ADJ)
CASUALTY (N)
CHARACTER (N)
COLLAR (N)
COLLEGE (N)
COMMODITY (N)
COMPREHENSIVE (ADJ)
CONDUCTOR (N)
CONFERENCE (N)
CONFIDENT (ADJ)
CRIME (N)
CRIMINAL (N)
DISCUSSION (N)
DIVERSION (N)
EMBARRASSED (ADJ)
ESTATE (N)
EVENTUALLY(ADV)
EXIT (N)
FABRIC (N)
FACILITIES (N)
FATAL (ADJ)
FIGURE (N)
FILE (N)
FINE (ADJ)
FIRM (N)
FRESH (ADJ)
INHABITED (ADJ)
LARGE (ADJ)
LECTURE (N)
LIBRARY (N)
LOCALS (N)
LUXURY (N)
MAYOR (N)
MOLEST (VB)
MOTORIST (N)
NOTE (N)
NOTICE (VB)
NOTICE (N)
OCCURRENCE (N)
OFFENCE (N)
OFFICE (N)
PAPER (N)
PARENTS (N)
PIPE (N)
PLATE (N)
POLICY (N)
PRACTICE (N)
PRACTISE (VB)
PRESERVATIVE (N)
PRESUME (VB)
PRETEND (VB)
PROFESSOR (N)
QUALIFICATIONS (N)
QUIET (ADJ)
RARE (ADJ)
REALISE (VB)
RECORD (N)
RECORD (VB)
REGULAR (ADJ)
REMOVE (VB)
RESUME (VB)
ROPE (N)
SENSIBLE (ADJ)
SOAP (N)
SOLICITOR (N)
STAMP (N)
STRANGER (N)
SUCCEED (VB)
SUCCESS (N)
SUPPORT (VB)
SYMPATHETIC (ADJ)
SYMPATHY (N)
TAP (N)
TOPIC (N)
ULTIMATE (ADJ)
ULTIMATELY (ADV)
URGE (V)
VARIOUS (ADJ)
VICIOUS (ADJ)

The search output provided by these two learner corpora was carefully examined by applying both quantitative and qualitative analyses so as to have a comprehensive view of *when and how* Spanish learners resort to English false friends in their written and spoken production. The sample sentences were manually analysed and sorted into correct and incorrect uses. The process was time-consuming and required good decision making to obtain reliable results. The analysis was complex since it involved registering frequencies and analysing the examples qualitatively. The data of learners' speech and writing were analysed separately so as to be able to compare the results in both modes of expression. Interpretative difficulties were related mainly to the ambiguous use of some false friends, which required me to try and understand what learners meant when they used certain items. In case of doubt, I consulted with native speakers to find out if the lexical items could be used in the way the students did and if these were natural uses in English.

4. Results

This section first discusses the quantitative results, i.e. the frequency of occurrence of each of the 100 false friends analysed. Then, it provides a more qualitative analysis, looking for explanations of the different percentages of inaccurate uses found for the individual lexical items and across the spoken and written data.

Table 2 below shows the total number of occurrences of each false friend, together with the percentages of incorrect and correct uses. Thus the columns labelled "raw frequency" show the frequency of occurrence of the 100 items under analysis in both ICLE and LINDSEI, and the columns to their right indicate the percentage of incorrect uses (% of inaccuracy).

Table 2. Frequencies and percentages in ICLE and LINDSEI

ENGLISH FALSE FRIENDS	ICLE		LINDSEI	
	RAW FREQUENCY	% of Inaccuracy	RAW FREQUENCY	% of Inaccuracy
Accommodate	0	0	0	0
Actual	19	89.5	13	7.7
Actually	21	47.6	41	2.4
Adequate	4	50	0	0
Advertise	5	20	0	0
Advice	4	50	1	0
Advise	1	0	0	0
Announce	2	100	0	0
Apparent	2	0	0	0
Appoint	1	100	0	0
Argument	8	37.5	5	0
Assist	4	75	0	0
Attend	18	38.9	2	0
Balance	13	7.7	0	0
Bank	12	0	0	0
Batteries	2	0	0	0
Bizarre	0	0	0	0
Blank	4	0	0	0
Camp	1	0	5	0
Career	14	85.7	10	100
Carpet	3	0	0	0
Casual	2	100	0	0
Casualty	5	40	0	0
Character	9	44.4	16	0

Collar	0	0	0	0
College	4	0	22	0
Commodity	2	100	0	0
Comprehensive	0	0	0	0
Conductor	0	0	0	0
Conference	0	0	0	0
Confident	4	0	3	0
Crime	134	1.5	1	0
Criminal	152	1.3	0	0
Discussion	20	5	0	0
Diversion	0	0	0	0
Embarrassed	2	0	2	0
Estate	1	100	0	0
Eventually	4	0	1	0
Exit	1	0	0	0
Fabric	0	0	0	0
Facilities	7	71.4	1	100
Fatal	0	0	0	0
Figure	65	16.9	2	0
File	1	0	0	0
Fine	5	0	10	0
Firm	5	0	1	0
Fresh	1	0	0	0
Inhabited	3	66.7	0	0
Large	23	43.5	1	100
Lecture	2	50	0	0
Library	4	0	6	0
Locals	0	0	0	0
Luxury	7	0	0	0
Mayor	0	0	0	0
Molest	0	0	0	0
Motorist	0	0	0	0

Note	0	0	1	0
Notice	4	75	9	0
Notice (verb)	19	5.3	0	0
Occurrence	0	0	0	0
Offence	13	0	0	0
Office	9	0	2	0
Paper	40	10	5	40
Parent(s)	41	0	30	0
Pipe	2	0	0	0
Plate	0	0	0	0
Policy	12	8.3	0	0
Practice (noun)	50	36	5	20
Practise (verb)	14	35.7	8	0
Preservative	0	0	0	0
Presume	1	100	1	100
Pretend	50	16	11	45.5
Professor	6	100	6	16.7
Qualifications	1	0	1	100
Quiet	5	60	8	50
Rare	1	100	1	100
Realise	88	8.0	18	0
Record (noun)	4	0	0	0
Record (verb)	2	0	0	0
Regular	2	0	2	50
Remove	5	0	0	0
Resume	3	100	0	0
Rope	1	0	0	0
Sensible	3	66.7	1	0
Soap	9	0	0	0
Solicitor	0	0	0	0
Stamp	0	0	0	0
Stranger	1	0	4	50

Succeed	4	0	0	0
Success	20	10	1	0
Support	42	11.9	0	0
Sympathetic	4	0	0	0
Sympathy	8	37.5	0	0
Tap	0	0	0	0
Topic	47	2.1	23	0
Ultimate	1	0	0	0
Ultimately	2	50	0	0
Urge	2	0	0	0
Various	4	50	0	0
Vicious	2	0	0	0

TOTAL	1,123	16.3	280	11.8

If we look at the overall results from both corpora, we see that a total of 1403 sample sentences were found to contain the false friends under examination, 1123 in ICLE and 280 in LINDSEI. Out of these 1403 examples, a total amount of 216 in the two corpora show an incorrect use of false friends. This volume of errors amounts to a percentage of 15.4% of the instances found (Figure 1).

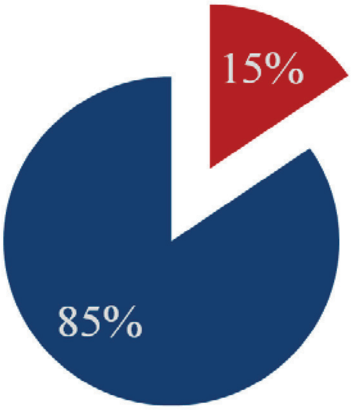


Figure 1. Overall results

Although 15 percent of the total may not be considered a high proportion of errors, it is quite significant since we are looking at high-frequency English words. Moreover, when these words are misused by learners, they may produce important misunderstandings. A more qualitative analysis confirms that the misuse of these lexical items may generate confusion and misinterpretations. The learners' use of words such as *resume, facilities* or *casualties* are good examples of the types of problems that Spanish learners may have with these lexical items.

(1) The play is, in **resume**, a critique of the absurdity of all forms and conventions. <ICLE-SP-UCM-0016.8>

(2) <B> yeah no there's a lot of creativity now in . that's it here in in Europe there are (mm) better ideas but we don't have the[i:] the[i:] **facilities** and the yes . they have the money they have (er) special effects and all that they have good stories also they have . best movies of history they are they are American but but Europe . Europe they have good good ideas </B><LINDSEI_SP049>

(3) […] miracles will be only a series of **casualties** or coincidences intelectually interpreted to the interest of a determined group (the Church in this case) <ICLE-SP-UCM-0007.6>

As shown in examples (1-3), Spanish learners use *in resume* with the intended meaning 'in short, to sum up', the plural noun *facilities* is found in a context where a word like 'opportunities' would be used by a native speaker and the noun *casualties* is used as if it were a synonym of 'coincidence'. Consequently, the receiver of the message may have difficulties understanding what the learner really means.

In a similar fashion, the uses of the English noun *career* reveal that learners tend to assign the Spanish sense of 'following a degree' to this word both in their spoken and written production, as illustrated by examples (4) and (5). The real sense of this English word, 'job or profession', is completely missing.

(4) When you choose to study an university **career**, you expect you may get a job within the branch you have chosen; but in the majority of the cases, that is not so. <ICLE-SP-UCM-0030.4>

(5) <B> yes . it's my . but it's my second (eh) **career** the one I'm doing now . and the first one I did from the teaching . training but infant education </B><LINDSEI_SP021>

Another high-frequency English word that deserves our attention here is the adjective *rare*. English *rare* has the sense 'uncommon, occurring infrequently'. However, Spanish learners of English seem to use it in the sense of 'strange, odd'.[3]

(6) religion alienation is doing a social function […] If we put into effect this concepts and ideas to the present they sound really **rare**, and they can even produce us laugh. <ICLE-SP-UCM-0051.3>

(7) <B> (uhu) and it's (er) very jo= it's like a joke because . I: go to France and <starts laughing> speak in English <stops laughing> it's: it's very **rare** </B><LINDSEI_SP047>

Problems with high-frequency English words such as *career* and *rare* show that learners draw strong associations between the lexical stocks of their mother tongue and their second language. As a consequence, despite the semantic differences and the frequency of these words in each language, whenever there are similarities between two lexical items, the L1 meaning seems to prevail and is attributed to the English lookalike.

Surprisingly, a comparison of the results gathered from ICLE and LINDSEI reveals that there are more errors in the learners' written than in their spoken production. However, if we take into account the overall frequencies in relation to the errors found in the two corpora, that these differences are not statistically significant (χ2: *p*-value= 0.1049, p < .0001).

Table 3. General summary of the data: ICLE *versus* LINDSEI

ICLE			**LINDSEI**		
RAW FREQUENCY	Incorrect Uses	% of Inaccuracy	RAW FREQUENCY	Incorrect Uses	% of Inaccuracy
1123	183	**16.3**	280	33	**11.8**

A possible explanation for this finding lies in the stylistic and register features of the language that learners use in the two modes of expression. Learners tend to use more basic vocabulary in their spoken production (LINDSEI) than in their written performance (ICLE). Thus, words such as *parents* or *topic* are common in spoken language. By contrast, if we look at the most high-frequently used words in written language, we find words such as *criminal, crime, realise, figure* and *practice*. It appears that learners use more straightforward and widely-used lexical items in the spoken mode, but look for more lexical variety and complexity in their written compositions.

Interestingly, the verb r*ealise* is a commonly used false friend in both ICLE and LINDSEI, but it exhibits a higher number of mistakes in the written corpus. The same goes for *actual, actually* and *notice* (noun). The language used in speech is more formulaic and simple (e.g. <B> *but I think . maybe she ha= (eh) .. ah look . I didn't* ***realise*** *. that she in this picture* […] </B> <LINDSEI_SP019>), and learners seem to display better idiomatic and collocational control in this mode. By contrast, in writing learners strive for a more sophisticated and varied vocabulary, which leads them to make more mistakes (e.g. *Soldiers may* ***realize*** *projects on different matters and also put them into effect into the barracks*. <ICLE-SP-UCM-0006.2>).This may explain why there are more errors in written production.

At any rate, the results of my analysis show that high-frequency English words which are false friends with the learners' mother tongue pose serious problems that have to be addressed in EFL classrooms. High-frequency words such as *career* or *rare* (among others) are frequently misused by learners in both spoken and written language. The fact that these 100 false friends are likely to be found in both receptive and productive processes does not help learners to acquire these lexical items accurately.

5. General conclusions

The main aim of this corpus-based study was to examine Spanish learners' use of false friends in speech and writing with a view to identifying the learners' problems with these lexical items so as to help teachers meet the learners' needs in these areas. Thus, a set of 100 carefully selected high-frequency false friends was investigated in two computerised learner corpora, ICLE and LINDSEI, which contain samples of written and spoken texts produced by Spanish learners of English.

In the 318,912 words analysed, there occurred 1,403 examples of the false friends considered in this study. Accordingly, there occur 2 of those 100 false friends every 500 words. In a general sense, the number of false friends that are accurately used is higher than the number that are incorrectly used. Nonetheless, the influence of the students' mother tongue is perceived in 15 per cent of the total. It is worth noting here that this study shows with empirical data that mistakes with English false friends still persist in the productions of advanced learners.

My analysis revealed that Spanish students deal in different ways with alleged typical false Not all of the 100 false friends under examination are equally problematic for learners in their L2 performance. Some of them are certainly more challenging than others. A noun, such as *career*, which is also quite common in learner language, is persistently causing problems. Learners use it to mean "university course", which constitutes an obvious case of crosslinguistic transfer in both speech and writing. In spite of the fact that this word is said to be continuously corrected by teachers in the classroom, the associations between English *career* and Spanish *carrera* appear to be so strong that it is impossible for students to keep these words separate in their mental lexicons. Students appear to be inevitably tempted to use this word in the wrong context when they are talking about their university studies (e.g. *Few years ago, the study of a* ***career*** *was destinated to the offsprings from wealthy families [...]* <ICLE-SP-UCM-0001.3>). In this case, teachers must mention this problem explicitly and provide learners with some clues so as to avoid any mistakes that could arise from the misuse of this word. By contrast, the word *parents* is the only one of the frequent words claimed to be false friends which shows no problems of semantic interference and seems to have been fully acquired. The concreteness of this noun, together with its early introduction in English courses, could have helped learners achieve a full command of it by this stage.

This study also showed that learners use more false friends in their written than in their spoken production. By the same token, the number of mistakes in written language is higher than in speech. Thus, the percentage of errors is 16 per cent in writing versus 12 per cent in speaking. At first glance, the data seem to indicate that the revision and editing actions which characterise the writing process are not effective when it comes to getting rid of these lexical errors. However, we have seen that learners make use of more complex vocabulary in their written than in their spoken performance which could explain the higher proportion of errors in writing.

On the whole, if we pay attention to the detailed analysis of the meanings learners attach to the 100 false friends under investigation, we see that, in a considerable amount of cases, learners are not acquainted with the semantic divergence that exist between certain English words and their mother tongue lookalikes. Accordingly, Spanish students use some English terms such as *actual* or *career* in the Spanish way instead of in the English way. However, this statement is too broad and simplistic and ignores individual differences between the items analysed. A one-by-one analysis of the 100 false friends examined shows that there are differences. Not all the false friends examined exhibit the same degree of semantic divergence and the same degree of difficulty. Although there are false friends, such as *career,* which seem to be very difficult for learners to use appropriately, there are others which have been appropriately acquired by learners, such as *embarrassed.* In addition to this, there are different types of false friends: there are total and partial false friends, i.e. false friends which mean totally different things in L1 and L2 and false friends which sometimes overlap semantically in both languages. This last group, the partial false friends, are less problematic (e.g. *crime*) as a result of the semantic coincidence between the two languages.

When we try to account for the reasons why learners use false friends wrongly, everything indicates that the origin of most problems is in the effect of crosslinguistic influence. Spanish learners use some English false friends as if they were true translation equivalents for their Spanish quasi-homograph counterparts. This excessive reliance on their mother tongue may be due to the students' inadequate level of English, to the bad quality of the input they receive or to the strategies they use to communicate in the L2. It is also likely that the learners' exposure to the foreign language has not been sufficient to erase these errors from their spoken and written productions.

We can conclude that what seems to be needed for the correct acquisition and use of these lexical items is: increased exposure of the learners to the English language, early incorporation of these lexical items in language learning, the use of suitable techniques for the teaching and recycling of these peculiar words (e.g. meaningful examples, clear contexts of use) - and the learners' efforts to learn these words and to be accurate in their use of English vocabulary. This leads us to a final discussion of some pedagogical implications.

6. Pedagogical implications

Learner corpora provide language experts with valuable information on the specific problems that learners have with false friends. Study of these databases may help teachers foresee and prevent this type of problems. Language teachers should not overlook this type of mistake since the misuse of false friends might distort the message of the speaker, as is the case in the following utterances: *Nowadays in the middle of this kind of life based in the* ***comodity*** *and the try to be in a high level of life[..]* <ICLE-SP-UCM-0041.3>; *In other places such as New York, where it never have snowed in this way, last week there were people incomunicated and they were* ***supporting*** *temperatures of fifteen degrees bellow zero.* <ICLE-SP-UCM-0052.4>).

Language instructors should raise the learners' awareness of false friends. Explicit instruction is needed (Chacón Beltrán 2006) to make learners aware of the semantic differences between similar items in the second language and their own language in a conscious way. Lengeling (1996: 5) suggests vocabulary teaching practices which might be useful for the effective teaching of false friends. She urges teachers to use strategies, such as *explaining* "how these words are different and what the correct word is for the corresponding word in the target language", *collecting* "those FF that cause problems and incorporate their teaching in the classroom" and *recycling* those problematic items from time to time.

In general, more attention should be paid to false friends in the English class. The teachers' output and an early introduction to meaning differences between the L1 and L2 might be effective. Thus, the learners' recurrent exposure to words such as *parents* and *topic* in the English classroom may explain the absence of mistakes in the use of these words. On the other hand, evidence from the two corpora suggests that teachers should not presume that well-known false friends which may have been studied already in earlier courses do not deserve attention. It may be the case that advanced students forget the semantic peculiarities of basic words (e.g. English *sensible* is misused in *Uncultivated people that are more* ***sensible*** *and accessible to external influences [...]*<ICLE-SP-UCM-0007.4>). Thus, students with an advanced level of the language also need information on English false friends. In addition to this, teachers should emphasise the role of context in vocabulary learning and should not approach lexis as a compilation of single words with fixed meanings. Audiovisual materials (e.g. pictures, audio files) and other teaching techniques, such as providing illustrative examples and suitable collocations, might be useful tools to promote the students' learning and correct use of English false friends.

In fact, the teachers' attitudes towards these words and their strategies to deal with them constitute a path for further research. It would be interesting to conduct a survey on the techniques and strategies used by language teachers to explain these lexical items in order to test the usefulness of these measures to reduce the lexical problem of false friends.

Notes

[1] I am grateful to the Spanish Ministry of Education (grant n. AP 2007-04477), to the European Regional Development Fund and the Autonomous Government of Galicia (Directorate General for Scientific and Technological Promotion, grant CN2011/011; CN2012/81) and to the Galician Ministry of Education (PGIDIT05PXIB20401PR and HU 2008/047) for their generous financial support.

[2] The *Longman Communication* 3000 *Word List* is based on the *390 million word-Longman Corpus Network,* a large database made up of different subcorpora, which is claimed to represent the core of English vocabulary. Kilgarriff´s frequency list is based on the *British National Corpus* (BNC). It contains 6,318 words which occur over 800 times in the 100 million-word corpus. The creation process of this list replicates the one used in the *Longman Dictionary of Contemporary English* (LDOCE) which makes it the best word list to complement the data provided by the *Longman Communication 3000 Word List.*

[3] The use of *rare* seems to be imbued with the negative sense of "weird", "surprising" or "unusual" in Spanish learner language. However, this English adjective does not necessarily have negative connotations as in *He plays with rare sensitivity,* which means that his way of playing is unusually good and remarkable.

References

Batchelor, R. and M. Offord (2000), *Using French: A Guide to Contemporary Usage* (3rd ed.). Cambridge: CUP.

Chacón Beltrán, R. (2006), 'Towards a typological classification of false friends (Spanish-English)', *Revista Española de Lingüística Aplicada*, 19: 29-39.

Cuenca Villarejo, M. (1987), *Diccionario de Términos Equívocos* ('Falsos Amigos') Inglés-Español-Inglés. Madrid: Editorial Alhambra.

Frantzen, D. (1998), 'Intrinsic and extrinsic factors that contribute to the difficulty of learning false cognates', *Foreign Language Annals*, 31: 243-254.

Frutos Martínez, M. C. (2001), 'El Problema de los Falsos Amigos en dos Lenguas Afines: Gallego e Italiano. Propuesta Metodológica', in: S. Porras Castro (ed.) *Lengua y Lenguaje poético: actas del IX Congreso Nacional de Italianistas*. Valladolid: Universidad de Valladolid. 287-294.

Gilquin, G., S. De Cock and S. Granger (2010), *Louvain International Database of Spoken English Interlanguage* (CD-ROM). Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. (1996), 'Words in English: From history to pedagogy', *KVHAA Konferenser*, 36: 105-121.

Granger, S., E. Dagneaux and F. Meunier (2002), *International Corpus of Learner English* (CD-ROM). Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. and H. Swallow (1988), ‘False Friends: a kaleidoscope of translation difficulties’, *Langage el l’homme* 23: 108-120.

Hill, R. (1982), *A Dictionary of False Friends.* London: Macmillan Press.

Kilgarriff, A. (2006), *BNC Lemmatised Word Frequency List*. Available online at http://www.kilgarriff.co.uk/BNClists/lemma.al.

Koessler, M. and J. Derocquigny (1928), *Les Faux Amis ou les Trahisons du Vocabulaire Anglais*. Paris: Librairie Vuibert.

Lado, R. (1957), *Linguistics Across Cultures.* Michigan: University of Michigan Press.

Malkiel, B. (2006), ‘The effect of translator training on interference and difficulty’, *Target*, 18: 337-366.

Palacios Martínez, I. and R. Alonso (2005), ‘Lexis and learner corpora: a study of English/Spanish false friends on the basis of the data provided by SULEC (Santiago University Learner of English Corpus)’, in: C. Mourón Figueroa and T. Moralejo Gárate (eds) *Proceedings of the 4th International Contrastive Linguistics Conference.* Santiago de Compostela: Servizo de Publicacións. 749-760.

Postigo Pinazo, E. (2007), *Diccionario de Falsos Amigos: Inglés-Español.* Madrid: Ediciones Verba.

Prado, M. (2001), *Diccionario de Falsos Amigos: Inglés-Español.* Madrid: Gredos, D.L.

Ruiz Mezcua, A. (2008), ‘Comparación de la cuestión de los falsos amigos en la interpretación y en la traducción’, in: R. Ramos Fernández and A. Ruiz Mezcua (eds) *Traducción y cultura. lenguas cercanas y lenguas lejanas: los falsos amigos.* Málaga: Encasa Ediciones, 155-171.

Sheen, R. (1997), ‘English *faux amis*/false friends for francophones learning English’, *Volterre-Fr English and French language resources*. Available online at http://www.wfi.fr/volterre/sheen.html.

Venuti, L. (2002), ‘The difference that translation makes: the translator’s unconscious’, in: A. Riccardi (eds) *Translation studies: perspectives on an emerging discipline*. Cambridge: CUP, 214-224.

Walsh, A. (2005), *False Friends and Semantic Shifts.* Granada: Universidad de Granada.

Automated classification of unexpected uses of *this* and *that* in a learner corpus of English

Thomas Gaillat, Pascale Sébillot** and Nicolas Ballier****

*Sorbonne Paris Cité and University of Rennes 1
**Institut de Recherche en Informatique et Systèmes Aléatoires (INRIA)
***Sorbonne Paris Cité

Abstract

This paper deals with the way learners make use of the demonstratives this *and* that*. NLP tools are applied to classify occurrences of native and non-native uses of the two forms. The objective of the two experiments is to automatically identify expected and unexpected uses. The textual environment of all the instances is explored at text and PoS level to uncover features which play a role in the selection of a particular form. Results of the first experiment show that the PoS features* predeterminer *and* determiner*, which are found in the immediate context, help identify unexpected learner uses among many occurrences, including native uses. The second experiment provides evidence that the PoS features* plural noun *and* coordinating conjunction *influence the unexpected uses of the demonstratives by learners. This study shows that NLP tools can be used to explore texts and uncover underlying grammatical categories that play a role in the selection of specific words.*

1. Introduction

In this paper, we present the results of two experiments designed to test automatic unexpected use classification of *this* and *that* in learner English. The objective is to analyse the textual environment of the demonstratives so as to uncover PoS or token-related features which play a role in the selection of a particular *expected* or *unexpected* form of *this* or *that*. This work is part of a wider project in which the objective is to annotate learner English, as it is clear that learner-error analysis requires such processed data (Granger 2008). We want to achieve this automatically by following de Haan (2000) using the ICLE corpus.[1] Our approach follows the path set by the Swiss linguist Frei (1929: 25), who insisted on the necessity to see errors as traces of language needs that have to be fulfilled by the learner. We endorse this view in which language facts are to be explained rather than just be compared with a norm. By using corpora and NLP tools as resources to validate linguistic explanations of particular language issues, we follow Frei's functional approach to linguistics.

Our work builds on previous work in several domains. Error tagging of learner corpora (Dagneaux et al. 1998, de Mönnink 2000) has shown that learner English requires specific processing, be it manual or computer-assisted. However, the increasing amount of data makes the manual annotation task a burden. In parallel, NLP tools have been playing an increasing role in annotation. Native corpora have been the target for automatic PoS and syntactic annotation for quite some time, with results showing accuracy of around 95% for the former type (Schmid 1994). It was not long until a learner corpus was also automatically annotated at PoS level (de Haan 2000, van Rooy and Schafer 2003). Recently, machine learning technologies have been applied to automatically detect various types of errors such as article selection (Han et al. 2006, Pradhan et al. 2010). Our approach follows the same line of research by using automatic PoS-tagging for automatic classification on a learner corpus. Instead of article selection, the focus is on the selection of demonstratives.

The objective of the paper is to describe two experiments carried out to discover the linguistic characteristics that influence the use of *this* and *that* in expected and unexpected contexts. The principle is to pass on a representation of contexts to a classifier in order to simulate the selection process of an instance of *this* or *that*. Our hypothesis is that the selection of *this* or *that* depends on its immediate context of utterance composed of PoS and words, called tokens in this article (Cornish 1999: 68). We build a representation of contexts by converting PoS and text into a set of features that a classifier arranges by classes according to metrics predetermined in a training phase. The first experiment is an attempt to measure the impact of certain distributional features of the demonstratives on their classification as *expected* or *unexpected* forms.[2] It discriminates *expected* uses of the two forms without distinction against *unexpected* uses of the same two forms. By selecting specific PoS tags from surrounding contexts of *expected* and *unexpected* occurrences, a classifying process is implemented to see whether or not these selected features play a role in the distinction between *expected* and *unexpected* uses. The second experiment's novelty lies in the nature of the dataset, as only *unexpected* uses of the demonstratives in their immediate context are considered. By using an automatic classifier on this dataset, the goal is to see which specific linguistic features play a role in the classification process of only the *unexpected* uses of *this* or *that*. In other words, the point is to see if the set of features in *unexpected* contexts helps to predict a particular *unexpected* form. Section 2 of the paper covers the method used to prepare the datasets. Section 3 deals with the way features are selected and extracted from texts. Finally, Section 4 discusses the results.

2. The dataset

In this part we describe the two components of the dataset and we explain how they are used in relation to the two experiments.

2.1 Native corpus subset

The dataset, made up of occurrences of *this* and *that,* comes from two corpora. A first subset consists of forty occurrences from the Penn Treebank-tagged Wall

Street Journal corpus (Charniak et al. 1987) which were extracted thanks to Tregex (Levy and Andrew 2006). The objective of the extraction process was to obtain twenty singular occurrences of each form, together with their surrounding context composed of token PoS tag pairs. The forty occurrences were selected randomly. The WSJ was chosen because of the quality of its PoS tag accuracy as the error rate is estimated at 3% (Marcus et al. 1993). A previous study involved the creation of new PoS tags for *this* and *that*, so as to provide more accuracy in the distinction of the forms.[3] The tagset includes a clear distinction between the determiner and pro-form functions of the demonstratives. The choice of this corpus also reflects the need to have a good reference point with the learner corpus. Michael Barlow (2005: 345) points out the problem of multiple genres in corpora: "the combination of genres in the general corpus does not provide a good reference point for the learner corpus, which invariably consists of a single genre". The WSJ provides a single genre with which other corpora may be compared. Even if it is a written corpus, single genre and reliable PoS annotation appear as strong factors for the choice of this corpus in the experiments.

2.2 Learner corpus subset

The second subset of the data is an extract from Charliphonia, the University of Paris-Diderot's subset of the Longdale corpus initiated by the Centre for English Corpus Linguistics at the University of Louvain.[4] This corpus is composed of audio recordings of English learners. Learners were interviewed by a native speaker and asked a few general questions on their recent background. Learners answered these questions exhaustively without many interruptions from the natives. For our experiments forty occurrences of *unexpected* uses of *this* and *that* were identified manually and extracted from the transcripts. Each occurrence was selected with its surrounding context. When the context included an occurrence of a demonstrative which was expected, the context was shortened so as to neutralise any expected use of the form. Without neutralisation, expected uses of the form would also be processed and, thus, introduce a bias to the homogeneity of the dataset. This sample includes 20 occurrences of *this* and 20 occurrences of *that*, which correspond to the two grammatical functions. As the sample is small, and in order to avoid variability due to number, only singular occurrences were selected. Consequently, only 40 occurrences of singular forms were selected randomly in the WSJ. The selection of *unexpected* uses was performed manually, and cross-validation was carried out with a native English speaker. A form was characterised as unexpected when the native speaker considered the choice of the demonstrative as not being the obvious one. A previous study shows that alternatives are substitutions with the other demonstrative or with the pronoun *it* or the determiner *the*.[5] In other words, unexpectedness is due to two trends: either the learners swap the two words or they swap the demonstrative with an erroneous use of *the* or *it*.

The following examples show the diversity of uses that may be classified as unexpected. In the first one, *this* and *that* are used as pro-forms and refer to the entity ‘pizza’. Native informers consulted on this example favour the pronoun *it* in both cases.

(1) DID0115-S001 “<A> would you consider pizza an Italian food </A> <B> (em) yes but it’s not it’s not really f= it’s typic but it’s not (em) we can eat **that** everyday everywhere now and . but (em) my grandma does **this** by herself”

In the second example, the demonstrative *that* is used in the position of determiner. Native speakers clearly favour the use of *the* or *this.* The choice of *that* creates a local context of rejection as if the country is of no interest to the speaker. The broader context proves the opposite as the speaker insists on her motivation to live in this country.

(2) DID0145-S002 “[I suppose I’ll go to Peru because this is a country I always (er) have intrigued me (er) my mummy my mom gave me a necklace with the God of Sun . Inti and since I had this like four five years ago I have always wanted to go there and to climb the Machu Picchu . so I will I want to go there so .] I’m gonna there for sure (em) . I will .. I will go there for a year I think (er) to work there to help people there . and to discover **that** country . because there are a lot a lot of things to: to find .”

(3) In the last example, **this** is used as a determiner. Unexpected use can even be classified as an error as the agreement between the form and its noun is not respected.

(4) DID0074-S001 “<B> (er) sports (em) not no sports but (em) music (em) because they they (em) in in **this** countries (em) (er)”

2.3 Corpus subsets for the experiments

For the first experiment, we use both subsets described above as we want to see what features lead to the distinction between learner corpus demonstratives, characterised by *unexpected* uses, and native-corpus demonstratives. This allows the identification of PoS and token elements that differentiate *expected* from *unexpected* uses. We do not distinguish between *this* or *that* at this point, but we introduce a balance number of the forms in the samples so that classification is not influenced by weight differences. Thus, the dataset for experiment 1 is composed of two subsets. The even number of occurrences of *this* and *that* forms in each subset gives a 50/50 baseline with which classification can be compared. The small size of the sample is due to the slow process of identifying *unexpected* uses manually. The classifying method explained below takes this into consideration so as to maximise training and test data.

The second experiment is an insight into *unexpected* learner English only. This is why, in a similar approach to Pradhan et al. (2010), the dataset is *only* composed of the Charliphonia subset described in Section 2.2. In other words, only *unexpected* uses are taken into consideration and the classification process is carried out so as to have a closer insight into the actual selection of a particular *unexpected* form. The idea is to identify what features lead to the selection of a particular unexpected form.

3. Features

In this section, we describe the way an abstract representation of the context is carried out so that the classifier can process the data. We show how the selection of features depends on linguistic criteria. The conversion of these criteria into features for the classifier and the classifying process itself end the section.

3.1 Selection of linguistic characteristics

For the purpose of our experiments, we needed to isolate the relevant characteristics in order to convert them into features for the classifier. All the literature on the subject identifies a number of notions that constitute characteristics in the uses of the forms. Biber et al. (1999: 347) distinguish the use of demonstratives according to the notion of distance: “In addition to marking something as known, the demonstrative forms specify whether the referent is near or distant in relation to the addressee”. Stirling (2002: 1504) endorses the same vision and adds a distinction between the dependent and independent uses of the demonstratives together with their deictic and anaphoric uses. Halliday and Hasan (1976: 56-68) encompass the same notions and integrate them within the system of endophoric and exophoric reference. Fraser and Joly (1979: 114) follow Hasan and Halliday in their vision of the system of reference, and go further in their analysis of the two forms. Anaphoric and deictic uses of the forms are distinguished within the reference system, and they introduce the notion of speaker’s sphere. The notions of conclusion, rejection, distantiation and rupture are usually accompanied with *that*, while the notions of speaker’s sphere, identification to the situation, proximity, personalisation, temporal location are usually found with *this.*

The challenge is thus to determine the PoS and words that could correspond to these characteristics depending on context. For both our experiments, we choose native English as a reference to establish a list of the characteristics to be tested. Even in the case of the learner corpus subset, we consider native-English characteristics as relevant since learners target the native language model when speaking. Table 1 shows the linguistic characteristics that have been selected according to their linguistic values and the PoS tags to which they correspond.

Table 1. Candidate features for expected uses and their linguistic justifications.

Description and characteristics	**Features (PoS tags*)**	**Context**
Verb in the preterite form in order to mark temporal distantiation within the context	VBD	Left
Punctuation in order to mark pauses and the speaker's attitude possibly signalling insistence or change of topic	PUNC*	Left + right
Personal pronouns in order to mark the speaker's existence	PRP	Left
Nouns in order to mark the possible co-reference of the demonstrative and the noun	NN	Left + right
Verb in order to mark predicate introduction, and thus of a reference to an entity	VB	Left + right
Wh- adverbs (where, when) and adverbs (not, never) that may mark the existence or rejection of detailed information on an entity	(W)RB	Left + right
Cardinal numbers such as pro-form *one* may mark the need to express contrast, i.e., "*this* one", "one of *these*"	CD	Left + right
Coordinating conjunction in order to mark possible changes of focus, reference or topic	CC	Left
Complementiser and relative pronoun that may mark the existence of detailed information on an entity	TCOM* / TREL*	Left
Determiner and pre-determiner in order to mark already existing determination of an entity	DT / PDT	Left
Modal in order to mark the speaker's possible attitude in relation to the situation of communication	MD	Right
Preposition in order to mark possible introduction to entity reference	IN	Left + right
Pro-form or determiner in order to mark the category of *this* or *that* for a particular instance	TPRON* / DT	Right

* Penn Treebank scheme except when there is an asterisk mark.

As far as words are concerned, they have been chosen according to several semantic groups that correspond to the notions identified above. Notions such as rejection (i.e., *no*, *never*), foreground/background information and interest (i.e., *want*, *hope*, *say*, *tell, first, second*), topic continuity/discontinuity (i.e., *after*, *however*, *then*) support the introspective choice of specific words. In addition, given the fact that the demonstratives are part of the domain of deixis, it was decided to include the words that provide referential information made by the speaker (i.e., *here, there, this, that, now*). The list of words also includes tokens expected to be found next to *this* or *that*, i.e.,*'s*, *all*, *lik*e, *of*.

As the second experiment only deals with unexpected uses in learner English, we think it is important to select specific linguistic characteristics for this experiment and add them to the aforementioned characteristics. Based on professional experience with learners, several characteristics are proposed. Experience in correcting both oral and written productions of students helped with the identification of grammatical issues that are found repeatedly amongst students. Firstly, the PoS tag giving information on the existence of plural nouns (NNS) in the right context is isolated, as it would help to determine agreement errors. Secondly, several words that are usually accompanied by learner difficulties are also isolated: *for, since, despite, (in) order, (in) spite* can all be part of direct translations from French, and as such, may appear in *unexpected* uses. The verb *is* is also isolated*,* as combinations with the demonstratives are not that clear for learners. In all these cases, learners make typical mistakes and the introduction of the words is an attempt to capture the environment in which errors with *this* or *that* occur. For example, some learners tend to say "In order this happen". By selecting the word *order* as a feature for the classifier, the idea is to see whether it helps with the improvement of the error classification process.

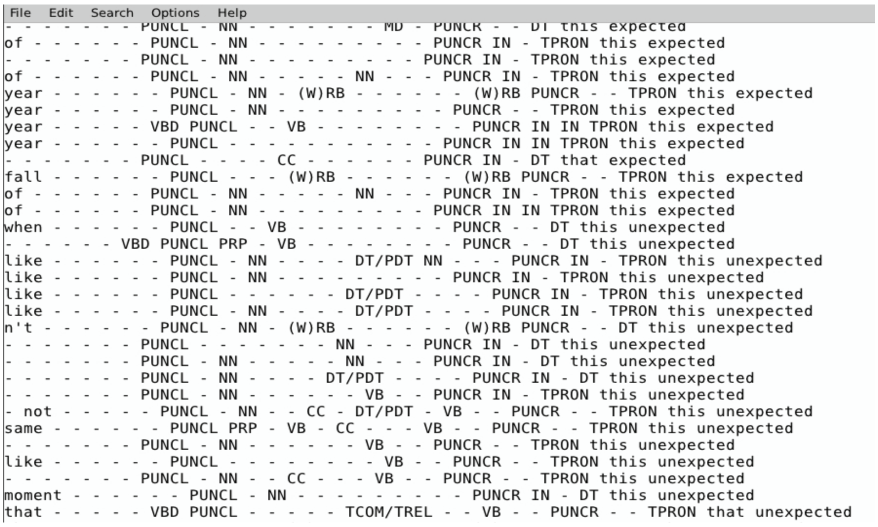
It is important to specify that no one feature can be seen as leading necessarily to the choice of a particular form. Instead, the experiment aims to test whether all the features, as a whole, have an influence or not on the choice of *this* or *that*. At this point in the study, it is not possible to indicate how the influence of feature x leads to the choice of *this* in one case, or *that* in another.

3.2 Extraction of an abstract feature representation

Before starting the classification process, the data containing the occurrences of *this* and *that* must be extracted so as to present a sequence of features to the classifier. The objective is to convert the previously mentioned characteristics present in texts into an abstract representation composed of lines of features. For each occurrence of the demonstratives, a sequencing PERL program scans the three preceding and following tokens and PoS tags to match them with the specific features expected to have an impact on the selection of demonstratives in native English. As a result, lines of features are created for each occurrence of the form and a class is assigned to each line of features.

For the first experiment, the program extracts features from the two subsets described in Section 2. This sequence of features is then matched to a particular class: *expected* or *unexpected*. For all the lines of features extracted from the native subset the class *expected* is assigned. For all the features extracted from the learner subset the class *unexpected* is assigned. Once the classes are assigned, both subsets are merged so as to finalise the training and test sets for the classifier. Figure 1 is a partial view of an extraction process where linguistic characteristics are turned into lines of features. The second line starts with the feature *of* as it was found three words before an occurrence of *this,* also printed as the second last feature of the same line. The hyphen sign after *of* means that none of the tokens listed in Section 3.1 were found two words before the occurrence of *this*. When PoS tags are matched by the PERL program they also are printed. The tags PUNCL and NN indicate that some punctuation and a noun were found within three words before the occurrence of *this*. A hyphen denotes a non-existing feature for a given position before or after the occurrence. The last element corresponds to the class assigned. When features are extracted from the native corpus subset described in Section 2.1, the *expected* class is printed like line two of Figure 1.

For the second experiment, a similar sequencing program is run on the Charliphonia subset described in Section 2.2 to create lines of features with their assigned class: *this* or *that*. In this experiment, extra features are scanned by the PERL program as the subset is characterised by the fact that it only includes *unexpected* uses of the forms. Since the objective of the experiment is to test features that lead to unexpected use, we have added non-native features based on the linguistics characteristics described in Section 3.1.



```
- - - - - - - PUNCL - NN - - - - - - - - MD - PUNCR - - DT this expected
of - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - TPRON this expected
- - - - - - - PUNCL - NN - - - - - - - - - - PUNCR IN - TPRON this expected
of - - - - - - PUNCL - NN - - - - - NN - - - PUNCR IN - TPRON this expected
year - - - - - - PUNCL - NN - (W)RB - - - - - - (W)RB PUNCR - - TPRON this expected
year - - - - - - PUNCL - NN - - - - - - - - - PUNCR - - TPRON this expected
year - - - - - VBD PUNCL - - VB - - - - - - - - PUNCR IN IN TPRON this expected
year - - - - - - PUNCL - - - - - - - - - - - PUNCR IN IN TPRON this expected
- - - - - - - PUNCL - - - - CC - - - - - - PUNCR IN - DT that expected
fall - - - - - - PUNCL - - - (W)RB - - - - - - (W)RB PUNCR - - TPRON this expected
of - - - - - - PUNCL - NN - - - - - NN - - - PUNCR IN - TPRON this expected
of - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN IN TPRON this expected
when - - - - - - PUNCL - - VB - - - - - - - - PUNCR - - DT this unexpected
- - - - - - VBD PUNCL PRP - VB - - - - - - - - PUNCR - - DT this unexpected
like - - - - - - PUNCL - NN - - - - DT/PDT NN - - - PUNCR IN - TPRON this unexpected
like - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - TPRON this unexpected
like - - - - - - PUNCL - - - - - - DT/PDT - - - - PUNCR IN - TPRON this unexpected
like - - - - - - PUNCL - NN - - - - DT/PDT - - - - PUNCR IN - TPRON this unexpected
n't - - - - - - PUNCL - NN - (W)RB - - - - - - (W)RB PUNCR - - DT this unexpected
- - - - - - - PUNCL - - - - - - - NN - - - PUNCR IN - DT this unexpected
- - - - - - - PUNCL - NN - - - - - NN - - - PUNCR IN - DT this unexpected
- - - - - - - PUNCL - NN - - - - DT/PDT - - - - PUNCR IN - DT this unexpected
- - - - - - - PUNCL - NN - - - - - - VB - - PUNCR IN - TPRON this unexpected
- not - - - - - PUNCL - NN - - CC - DT/PDT - VB - - PUNCR - - TPRON this unexpected
same - - - - - - PUNCL PRP - VB - CC - - - VB - - PUNCR - - TPRON this unexpected
- - - - - - - PUNCL - NN - - - - - - VB - - PUNCR - - TPRON this unexpected
like - - - - - - PUNCL - - - - - - - - - VB - - PUNCR - - TPRON this unexpected
- - - - - - - PUNCL - NN - - CC - - - VB - - PUNCR - - TPRON this unexpected
moment - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT this unexpected
that - - - - - VBD PUNCL - - - - - TCOM/TREL - - VB - - PUNCR - - TPRON that unexpected
```

Figure 1. Feature set for *expected* and *unexpected* classes

4. Classification and results

In this section, we explain the classifying method used by the classifier and how its performance is assessed. The second part deals with the results of the classification experiments.

4.1 Classification method

The machine-learning method used for the experiment applies the memory-based method, and the IB1 or k-nearest neighbour algorithm is implemented in TiMBL (Daelemans et al. 2010). In machine learning, two types of data are necessary in order to classify and verify its performance. The memory-based learner TiMBL first goes through a training phase before performing the classifying phase. In the training phase, it adds lines of features and their class to its memory. Each line constitutes a vector of features. In the classifying phase, the classifier predicts the class of new lines of features without the class information. The similarity between the new lines of features and all the examples in memory is computed using a distance metric. The prediction is made by assigning the most frequent category within the found set of most similar line(s), *i.e.,* the k lines memorised in the training phase that are nearest to the line being processed. To do so, the classifier computes a series of metrics (gain ratio) in order to establish the order of the features to be taken into account in the decision process. It establishes a hierarchy of the features from most relevant to least relevant in the classifying process.

Due to the low volume of data, we use the leave-one-out option for training and testing on our dataset, which means that for each instance of the experiment, only one line of the file is used for testing and the other patterns are used for training. This process is repeated for each pattern and the advantage is that, considering the small size of the samples, the leave-one-out option allows for greater robustness and generality to the learned hypothesis. "No test file is read, but testing is done on each pattern of the training file, by treating each pattern of the training file in turn as a test case (and the whole remainder of the file as training cases)." (Daelemans et al. 2010: 41). In order to evaluate the performance of the classification, precision and recall are calculated for each line due to the leave-one-out option. The results presented in Section 4.2 represent averages of each metric for successive classifying tests.

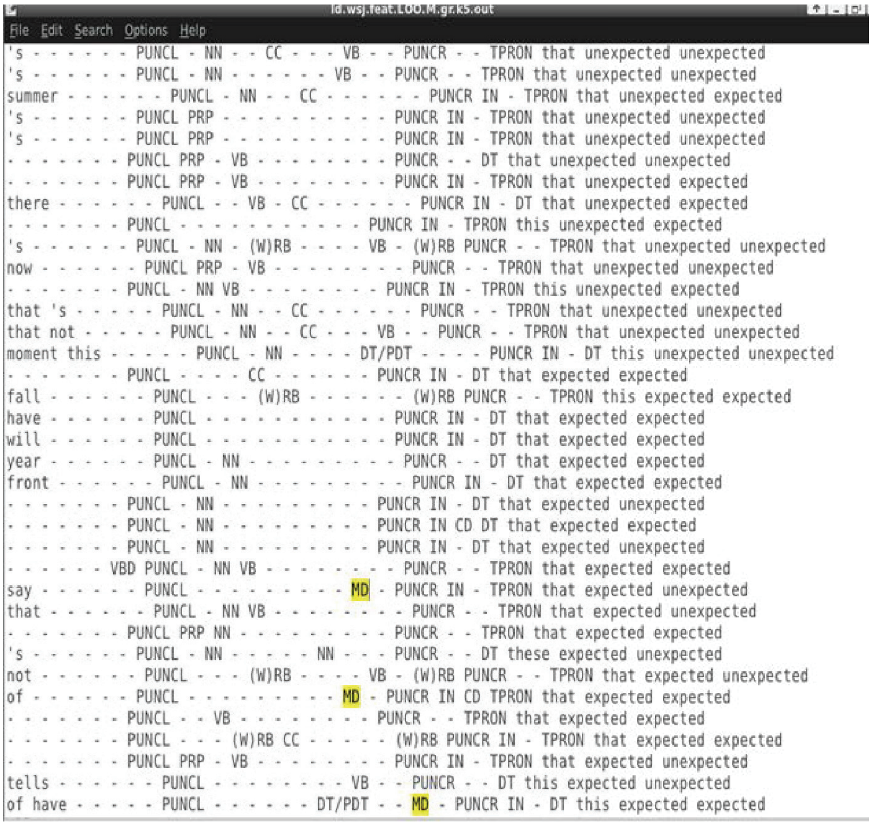
4.2 Results

We present the results of our experiments in two parts. First, we show the results of *unexpected* and *expected* classification in Table 2. The subsets used for experiment one include an equal number of *expected* and *unexpected* lines and an equal number of *this* and *that* occurrences. This means that random classification provides overall accuracy of 50%. Considering this 50/50 baseline of the subsets, the extra 20% improvement margin (the actual accuracy less the random accuracy) gives a measurement of the relevance of the feature set for the selection of expected *this* or *that*. This level remains rather low as NLP classification tasks usually show results well above 90%. The fact that not all lines obtain the correct class may be explained by a lack of exhaustiveness in the type of features. The immediate context of each occurrence, that is three tokens and three PoS tags before and after may be seen as a limitation as there may be linguistic characteristics located further away that influence the selection of a class. Another limitation may find its source in the difference between the oral and written modes of the native and non-native subsets. The mode of the WSJ is written while that of the non-native subset is oral. A bias may have been introduced due to differences linked to distinct style and syntactic-complexity profiles. Classification between *expected* and *unexpected* uses determines the extent to which the features have an impact on the selection of the forms.

Table 2. Experiment 1 – *Unexpected* and *expected* classification results

Scores per value class	precision	recall
expected	0.69048	0.72500
unexpected	0.72500	0.69048
overall accuracy: 0.707317		

The mixed subset approach shows that it is possible to distinguish between *unexpected* and *expected* forms thanks to the selection of particular features that the classifier uses to categorise the abstraction of occurrences. TiMBL allows the user to have access to the feature order set during the training phase. The gain ratio weight calculated for each feature shows the significance of each feature in the classification. Incidentally, it provides the linguist with significant information on each linguistic characteristic that is abstracted in the feature vectors. For experiment one, the first four features have a gain ratio above 10%. They are CD (Number) within three PoS tags to the left or right, and MD (Modal), TCOM/TREL (*that* as complementiser or relative pronoun) and DT/PDT (Determiner or pre-determiner) within 3 PoS tags to the left. For the classifier, the presence of a cardinal number in the left or right context is a prime criterion to differentiate between *expected* and *unexpected* uses of any demonstrative used by learners. If one looks at the data, it appears that CD is only used in the *expected* subset. Hence, it is logical that any new line including the CD feature is classified as *expected*. The TCOM/TREL feature only appears once in the data and so it also becomes a determining factor for classification in logical terms. It may be more appropriate to search for this tag in the right context of the forms as it can be argued that hypotaxis occurs to provide details of an entity expressed in its preceding NP. This search may have led to more occurrences of this feature, giving it more relevance in linguistic terms. The relevance of the CD and TCOM/TREL features, thus, may be questioned linguistically as their higher gain ratio may only be due to the data representation of the sample. On the other hand, the distribution of DT/PDT shows a different pattern in the data as it is found in many lines, but not all, for both classes. Classification shows that when the feature appears on a line, it leads to *unexpected* in 9 cases out of 11. Thus, while this feature is present across the data, the classification results suggest that it plays a significant role in helping the classifier differentiate an *expected* use of a demonstrative from an *unexpected* use. The data suggests it might be a feature of unexpectedness. The MD feature and its influence remains unclear. In the training data, it appears in *expected* uses only so it is logically found in lines classified as *expected*. However, it appears on one line classified as *unexpected* (For all comments see Figure 2).



```
's - - - - - - PUNCL - NN - - CC - - - VB - - PUNCR - - TPRON that unexpected unexpected
's - - - - - - PUNCL - NN - - - - - - VB - - PUNCR - - TPRON that unexpected unexpected
summer - - - - - - PUNCL - NN - - CC - - - - - - PUNCR IN - TPRON that unexpected expected
's - - - - - - PUNCL PRP - - - - - - - - - - PUNCR IN - TPRON that unexpected unexpected
's - - - - - - PUNCL PRP - - - - - - - - - - PUNCR IN - TPRON that unexpected unexpected
- - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR - - DT that unexpected unexpected
- - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR IN - TPRON that unexpected expected
there - - - - - - PUNCL - - VB - CC - - - - - - PUNCR IN - DT that unexpected expected
- - - - - - - PUNCL - - - - - - - - - - - PUNCR IN - TPRON this unexpected expected
's - - - - - - PUNCL - NN - (W)RB - - - - VB - (W)RB PUNCR - - TPRON that unexpected unexpected
now - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR - - TPRON that unexpected unexpected
- - - - - - - PUNCL - NN VB - - - - - - - - PUNCR IN - TPRON this unexpected expected
that 's - - - - - PUNCL - NN - - CC - - - - - - PUNCR - - TPRON that unexpected unexpected
that not - - - - - PUNCL - NN - - CC - - - VB - - PUNCR - - TPRON that unexpected unexpected
moment this - - - - - PUNCL - NN - - - - DT/PDT - - - - PUNCR IN - DT this unexpected unexpected
- - - - - - - PUNCL - - - - CC - - - - - - PUNCR IN - DT that expected expected
fall - - - - - - PUNCL - - - (W)RB - - - - - - (W)RB PUNCR - - TPRON this expected expected
have - - - - - - PUNCL - - - - - - - - - - - PUNCR IN - DT that expected expected
will - - - - - - PUNCL - - - - - - - - - - - PUNCR IN - DT that expected expected
year - - - - - - PUNCL - NN - - - - - - - - - PUNCR - - DT that expected expected
front - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT that expected expected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT that expected unexpected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN CD DT that expected expected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT that expected unexpected
- - - - - - VBD PUNCL - NN VB - - - - - - - - PUNCR - - TPRON that expected expected
say - - - - - - PUNCL - - - - - - - - - MD - PUNCR IN - TPRON that expected unexpected
that - - - - - - PUNCL - NN VB - - - - - - - - PUNCR - - TPRON that expected unexpected
- - - - - - - PUNCL PRP NN - - - - - - - - - PUNCR - - TPRON that expected expected
's - - - - - - PUNCL - NN - - - - - NN - - - PUNCR - - DT these expected unexpected
not - - - - - - PUNCL - - - (W)RB - - - - VB - (W)RB PUNCR - - TPRON that expected unexpected
of - - - - - - PUNCL - - - - - - - - - MD - PUNCR IN CD TPRON that expected expected
- - - - - - - PUNCL - - VB - - - - - - - - PUNCR - - TPRON that expected expected
- - - - - - - PUNCL - - - (W)RB CC - - - - - (W)RB PUNCR IN - TPRON that expected expected
- - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR IN - TPRON that expected unexpected
tells - - - - - - PUNCL - - - - - - - - VB - - PUNCR - - DT this expected unexpected
of have - - - - - PUNCL - - - - - - DT/PDT - - MD - PUNCR IN - DT this expected expected
```

Figure 2. Lines of features with their initial and automatically assigned classes

Experiment two gives information on learner specific features that lead to *unexpected* uses. This is why the second experiment is based solely on *unexpected* uses of learners, as it is intended to give more insight into the selection process of *unexpected* forms. The following results (Table 3) are in relation to the classification of *unexpected* forms only. There are two phases. Firstly, the experiment is carried out with features only selected for native English, or in other words, the features used in the first experiment. Secondly, the learner features mentioned above are added. For example, if we consider the column "overall accuracy", we obtain 0.80 accuracy when the features are extracted with non-learner specific features, and 0.88 when they are extracted with learner-specific features. The gain in accuracy is substantial, and shows that these features have an impact on the selection of *unexpected* forms. Even if more features related to learner use need to be tested, the process shows that it is possible to validate learner-related features that lead to *unexpected* uses. It also shows that features based on native English also partake in the *unexpected* form selection.

If one studies the order of the feature weights calculated by the classifier with non-learner specific features, the following features appear first: TCOM/TREL (*that* complementiser or relative pronoun), IN (Preposition), CC (Coordinating conjunction). The same calculation with learner-specific features gives the following order: NNS (plural noun), TCOM/TREL *(that* complementiser, relative pronoun), IN (Preposition) and CC (Coordinating conjunction). Thus, the way to distinguish learner unexpected uses of *this* from uses of *that* is done primarily *via* a feature denoting plural agreement error and it has a significant impact on overall accuracy. When observing the data, it appears that NNS is always linked to the unexpected selection of *this*. The following example contains the word *countries* in context:

(5) Speaker A: I haven't class this day er
Speaker B: sports
Speaker A: em not no sports but em music em because they they em in in **this countries** em er em movement er musical movement was born in the nineteenth er twentieth century

The fact that *countries* is PoS tagged as a plural noun with NNS, and that NNS is passed on to the classifier as a feature, teaches the classifier that the sequence *this* + NNS is not possible. As a consequence, any new occurrence of the sequence in any context is classified as *unexpected*.

The second group of features remains the same as with the non-learner specific extraction. In 9 cases out of 10, CC is related to the correctly assigned *that* class, making it a candidate for influencing the unexpected selection of *that*. Conversely, IN corresponds to 11 cases of correctly assigned *this* as opposed to 4 cases of correctly assigned *that,* which makes it a candidate for influencing the unexpected selection of *this*. To finish, TCOM/TREL appears only once in the data, which makes it a logical but not linguistically relevant factor.

Table 3. Experiment 2 – Classification of *unexpected this* and *that*, with and without learner specific features

Scores value class:	overall accuracy		precision		recall	
	Non learner specific	Learner specific features	Non learner specific	Learner specific features	Non learner specific	Learner specific features
this	N/A	N/A	0.77273	0.85714	0.85000	0.90000
that	N/A	N/A	0.85000	0.90476	0.77273	0.86364
	0.809524	0.880952				

5. Conclusion

In this article, we have covered the issue of automatic classification of learner English occurrences of unexpected uses of *this* and *that*. We have evaluated two types of classifications based on native and learner English. The first objective was to see how *unexpected* learner use of *this* and *that* could be predicted in cases including native and non-native use of the forms. The second objective was to uncover elements that may influence the selection of *unexpected* forms by learners. Answers to this second question would provide valuable information to support teaching to ESL students as teachers could make their students aware of specific recurrent features of unexpected uses of the forms. The approach adopted in this study allowed the contrastive exploration of non-native speech and native speech with the aim of finding features that influence linguistic choices. In line with Frei's view, errors are used as traces of language needs to be explained rather than to be compared with a norm.

Two types of data were used. In the first experiment, the classifier was trained with two equal subsets composed of native and non-native corpora. The non-native subset consisted of learner uses of *this* and *that* in an *unexpected* manner. Special care was given to the selection of features so as to make them correspond to linguistic notions developed in the literature on *this* and *that*. The classifying process was a test to distinguish between *expected* and *unexpected* forms, and results showed a 70% accuracy. The presence of determiners or predeterminers in the immediate context appeared as a significant feature for unexpectedness.

The second experiment was carried out to have a better insight into the selection of *unexpected* forms by learners. Several features were tested to measure the extent of their importance in the selection process. The feature related to plural nouns, i.e. a plural noun after a singular form of *this* or *that*, proved to have a substantial impact on the classification as it improved the performance by 7% compared with a classification based on features identified on native English. Overall, features such as *plural noun* and *preposition* in the previous context seemed to be factors for unexpected *this.* The *coordinating conjunction* feature may be a factor for the unexpected selection of *that.*

Future work includes the refinement of the feature selection process, as the one based on native English still needs more accurate and relevant information. For example, each occurrence of *this* and *that* could be annotated as being deictic or anaphoric. This kind of feature might help to identify contexts and their type of reference. More learner specific features also need to be identified and tested to explore further the way learners make their choices while speaking. These features also need to be tested on other non-native corpora including NOCE, a spoken corpus of learner English (Diaz Negrillo 2009). After identifying features of use for the demonstratives both in native and non-native speech, it will be possible to automatically introduce them in an annotation layer. As a result, the full process of detection and annotation will be automated. Queries on the demonstratives may then be launched on several corpora simultaneously. The ultimate objective is to carry out comparative analysis on the use of demonstratives, between learners of different L1s, or between natives and learners. This classification method could also be applied to other linguistic items. The process would require the selection of linguistic contextual features for the given item. Entire corpora could thus be processed to classify all occurrences of the item according to relevant categories for the analysis of this item.

Learner English is a fast-growing field of study and has been accompanied by the development of many corpora. Each corpus comes with its own meta-structure which makes cross-corpora querying impossible. By focusing on a particular linguistic point, at the end of our project it will be possible to develop a fine-grained automatic annotation process that will be applicable to any corpus. The final objective is to make it possible to import and query several corpora at the same time in order to carry out contrastive analyses depending on the nature of these corpora (native v. non-native, different L1s).

Notes

[1] The *International Corpus of Learner English* whose director is Sylviane Granger. http://www.uclouvain.be/en-cecl-icle.html.

[2] The term *unexpected* was favoured over the term error after tests on natives. Non-native occurrences were shown to natives. The tests consisted in presenting actual non-native utterances to natives with gaps replacing *this* and *that*. Natives were first asked to fill the gaps. When their choice contradicted the non-native choice, they were asked to judge the non-native choice. The tests showed that natives would classify choices in three categories: acceptable, unacceptable and acceptable as a second choice. The term *unexpected* covers both the unacceptable and second-choice categories.

3 This study focused on the detection of features that lead to the selection of demonstratives in their pro-form use in native English. It was done in collaboration with Detmar Meurers (University of Tubingen) and Nicolas Ballier (University of Paris-Diderot).

4 http://www.uclouvain.be/en-cecl-longdale.html.

5 In a paper (Gaillat 2013), presented at the conference "Learner Corpus Research 2011" in Louvain, I showed that, in learner use, interferences exist within the determination system as the demonstratives compete with the article *the*. Interferences within the anaphoric system also exist as demonstratives compete with the pronoun *it.*

Source

Charniak, E., D. Blaheta, N. Ge, K. Hall, J. Hale and M. Johnson (1987), 'BLLIP 1987-1989 WSJ Corpus Release 1'.

References

Barlow, Michael. 2005. 'Computer-based analyses of learner corpora', in: R. Ellis and G. Barkhuizen (eds) *Analysing Learner Language*. Oxford: OUP. 337-357.

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

Cornish, F. (1999), *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford: OUP.

Daelemans, W., J. Zavrel, K. van der Sloot and A. van den Bosch (2010), *TiMBL: Tilburg Memory-Based Learner Version 6.3 Reference Guide*. Tilburg, The Netherlands: Induction of Linguistic Knowledge, Tilburg University and CLiPS, University of Antwerp, Available online at http://ilk.uvt.nl/downloads/pub/papers/ilk.1001.pdf (last accessed on June 17, 2013).

Dagneaux, E., S. Denness and S. Granger (1998), 'Computer-aided error analysis', *System,* 26: 163-174.

Díaz Negrillo, A. (2009), *EARS: A User's Manual*. Munich: LINCOM Academic Reference Books.

Fraser, T. and A. Joly (1979), 'Le système de la deixis - Esquisse d'une théorie d'expression en anglais', *Modèles linguistiques,* 1: 97-157.

Frei, H. [1929] (2011), *La Grammaire des Fautes*. Rennes: Presses Universitaires de Rennes.

Gaillat, T. (2013), 'Towards a fine-grained annotation of *this* and *that*: a typology of use in native and learner English', in: S. Granger, G. Gilquin and F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. (2008), 'Learner corpora in foreign language education', in: *Encyclopedia of Language and Education*, 4: 337-351.

de Haan, P. (2000), 'Tagging non-native English with the TOSCA-ICLE tagger', in: C. Mair and M. Hundt (eds) *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi. 69-80.

Halliday, M.A.K. and R. Hasan (1976), *Cohesion in English*. London: Longman.

Han, N. R., M. Chodorow and C. Leacock (2006), 'Detecting errors in English article usage by non-native speakers', *Natural Language Engineering,* 2: 115-129.

Levy, R. and G. Andrew (2006), 'Tregex and Tsurgeon: tools for querying and manipulating tree data structures', *Proceedings of the 5th International Conference on Language Resources and Evaluation.* Available online at http://nlp.stanford.edu/pubs/levy_andrew_lrec2006.pdf (last accessed on June 17, 2013).

Marcus, M. P., M. A. Marcinkiewicz and B. Santorini (1993), 'Building a large annotated corpus of English: The Penn Treebank', *Computational Linguistics,* 19: 313-330.

de Mönnink, I. (2000), 'Parsing a learner corpus', in: C. Mair and M. Hundt (eds) *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi. 82-90.

Pradhan, A. M., A. S. Varde, J. Peng and E. M. Fitzpatrick (2010), 'Automatic cassification of article errors in L2 Written English', in: *Proceedings of the Twenty-Third International FLAIRS Conference*. Association for the Advancement of Artificial Intelligence (AAAI). Available online at https://www.aaai.org/ocs/index.php/FLAIRS/2010/paper/view/1342/1751 (last accessed on June 17, 2013).

van Rooy, B. and L. Schafer (2003), 'An evaluation of three PoS taggers for the tagging of the Tswana Learner English Corpus', in: *Proceedings of the Corpus Linguistics 2003 Conference.* Available online at http://www.corpus4u.org/upload/forum/2005092023174960.pdf (last accessed on June 17, 2013).

Schmid, H. (1994), 'Probabilistic Part-of-Speech tagging using decision trees', in: *Proceedings of the International Conference on New Methods in Language Processing.* Available online at http://www.stttelkom.ac.id/staf/imd/Riset/POS%20Tagging/Using%20Decision%20Tree.pdf (last accessed on June 17, 2013).

Stirling, L. (2002), 'Deixis and anaphora', in: R. Huddleston and G. K. Pullum (eds) *The Cambridge Grammar of the English Language.* Cambridge: CUP. 1449-1564.

Crude contours: a pilot study into the feasibility of charting student speakers' proficiency

†*Monique van der Haagen, Pieter de Haan and Rina de Vries*

Radboud University Nijmegen

Abstract

Dutch students of English at Radboud University, Nijmegen, the Netherlands are believed to enter university at CEFR B2 and, on graduation, are expected to have reached CEFR C2 in reading, writing, listening, spoken production and interaction. There is, however, preciously little evidence that links students' proficiency to the actual CEFR. This is hardly surprising as the difficulties of linking language users' production to specific CEFR levels are well-known. The English department does not really systematically chart students' progress in language proficiency, so that no documentation of the development of students' language production is available.

As a first attempt at finding out whether or not it is possible to measure students' progress in spoken English over the first two years of their degree course objectively, a small pilot study of a corpus of spoken English was undertaken. The participants were 31 students from a single cohort who were recorded in their first week at university and at the end of their first and second year. On all occasions they were asked to respond to a set of written general questions, such as "have you read a good book lately?".

1. Introduction

The department of English Language and Culture at Radboud University accepts students that either hold a secondary school diploma at *VWO* (*Voorbereidend Wetenschappelijk Onderwijs* – 'Preparatory Academic Education') level, i.e. pre-university education, broadly comparable to British A-levels, or a propaedeutic (= first-year) diploma obtained at an institute of higher professional education, usually at a teacher training college. Admission to such an institute requires a secondary school diploma at *HAVO*-level (*Hoger Algemeen Voortgezet Onderwijs* – 'Higher General Advanced Education'), i.e. senior general secondary education.

The English proficiency of these two groups of incoming students shows some discrepancy. The *Europees Referentiekader Talen* ('Common European Framework of Reference for Languages') website[1] identifies the CEFR levels that *VWO* and *HAVO* students should have reached for each of the skills at the end of their studies as follows:

Table 1. EFL skills defined for Dutch secondary school leaving examination

	HAVO	*VWO*
Listening	B1	B2
Writing	B1	B2
Spoken production	B1 (top end)	B2
Spoken interaction	B1 (top end)	B2
Reading	B1 (70% of questions at final exam at B2)	B2 (15% of questions at final exam at C1)

HAVO-graduates will obviously work on their English proficiency during their first year at a teacher training college, but it is unclear whether their proficiency will actually improve to B2 across the board within that year. As the English department at Radboud University does not hold entry exams and takes the incoming students' diplomas at face value, there is no way of knowing whether students are actually at B1 or B2 or what the gap, if any, might be between the two groups of students. Anecdotal evidence (exam results, lecturers' observations) suggests that such a gap not only exists, but is in fact considerable. What is more, the drop-out rate among the students that come to us without a *VWO* diploma is considerably higher than that of those with a *WVO* diploma, which is not seldom due to the fact that they fail to improve their English proficiency. This suggests that the year spent at the teacher training college has not been enough for them to catch up with the other students.

One of the aims of the English department is to educate our students to a near-native level of proficiency, specified on the departmental website as CEFR C2. No distinction is made between the five skills; graduates from Radboud University are expected to be at C2 level across the board.

In the first three years of study, they take seven proficiency courses, amounting to 252 hours of formal instruction. Of these, 140 are spent on oral production and pronunciation. However, as all departmental courses are taught in English and any reading, writing or presenting is in English, exposure to and use of the English language outside the proficiency classroom will contribute significantly to their level of proficiency in all the skills. In all, the average English graduate will have had 776 contact hours in the major programme (120 ECTS), during which English is either the vehicle of teaching or the subject of teaching.

Any attempt at identifying the number of hours of formal instruction required to get from one CEFR level to the next is to be regarded with some suspicion, but if we compare the total numbers of hours suggested in Pearson/Longman's *Teacher's guide to the Common European Framework*[2] with the minimum number of hours of instruction our students will have had under their belts at the end of their studies, it would seem that at 750-900 hours of formal instruction they may not meet the 1,000-1,200 estimate by Pearson/ Longman, but if we also take into the consideration class time spent in and on the

English language without specifically focusing on language acquisition, at 1,450 – 1,700 hours they are well over the estimated number of hours.

There is no doubt that the language skills of students improve significantly in the course of their studies (de Haan and van der Haagen 2012, 2013a), but any link to the CEFR is tenuous at best. In view of the fact that most other English departments in Dutch universities see their graduates as reaching CEFR C1, it is important for Radboud to chart the proficiency development of its students and relate this development to the proficiency teaching programme.

Earlier studies on advanced Dutch written EFL material have shown that development is not easy to determine on the basis of crude measures like type-token ratio, lexical density or lexical profiles (de Haan and van Esch 2005, 2008). Any development must be found in far more subtle lexical features, like appropriate collocations (de Haan and van der Haagen 2013b), or syntactic features, like information structuring (Verheijen, de Haan, and Los 2013). However, we hypothesise that since our EFL students produced unplanned speech, they may not be concerned with the finer details of vocabulary use and syntactic structures that they are probably aware of when they are planning (and possibly also editing) their written work. This is – at least for lexis – confirmed by Crookes (1989), who finds greater lexical variation in planned ESL speech than in unplanned ESL speech.

A first hurdle to overcome is to find out whether it is actually possible to obtain some hard, objective facts about students' spoken production development. We were fortunate enough to be able to run a pilot project on the basis of a small corpus (35,055 words) of unprepared speech produced in the context of pronunciation exams.

It was decided to use this corpus as a test to see if any development in students' spoken English over a period of two years could be made visible by running a number of some fairly crude tests, which were to yield data about speech rate, mean sentence length, type-token-ratio, lexical density and lexical sophistication. We know from research into advanced written learner English that such measurements are not helpful as the issues there are closely related to sentence construction, (de Haan and van Esch 2005, 2008), but we believe that such measurements might shed some light on spoken production. We therefore want to find an answer to the following research question:

> Can we chart the development of EFL student speakers' oral production on the basis of crude measures like speech rate, type-token ratio, word length, lexical density, and lexical sophistication?

This paper reports on the results of this pilot project.

2. Method

We analysed 93 recordings of unprepared speech made by 31 students at three different times. They were all students who started their studies in September 2006 (entrance level, henceforth referred to as T1), and who also made similar recordings at the end of their first year in June 2007 (T2) and at the end of their second year in June 2008 (T3). Their ages ranged from 18 to 21, the majority coming straight from secondary school. The students selected for this study were those that did not drop out during the first two years, and who took part in these three recording sessions. This does not necessarily mean that they were the best students (some of them did not pass the first year pronunciation and fluency exams in one go), but they can be assumed to be highly motivated learners.

During the test, the students first read out a short English text and then responded to a number of questions such as “How did you spend your summer holiday?” (T1), “How did you like your first year of English?” (T2), and “How does the second year compare to the first?” (T3). The students had an opportunity to spend a few minutes reading the questions before the recording began, but they were not allowed to make notes or write down complete answers. Students could not stop the recording, so any self-correction would have to be done instantaneously. The purpose of the pronunciation test is to collect both scripted and spontaneous speech, but for our data analysis we only collected the answers to the questions.

The data were analysed for speech rate (words/second), mean word length, type-token ratio, lexical density and lexical sophistication. These measures were chosen because they can arguably be considered indicative of general fluency and of syntactic and lexical sophistication (For speech rate, cf. Hincks 2010; Trofimovich and Baker 2006; for lexical density and lexical profiles in EFL writing, cf. Vidakovic and Barker 2009; Vidakovic and Barker 2010). It should be noted that lexical density may be a problematic measure. Johansson (2008) finds that it is a less effective measure for distinguishing between less mature and more mature Swedish L1 speakers than lexical diversity. We will come back to this in the discussion.

3. Results

Within-subject analyses of the data show that the speech rate and mean word length increased over time; however, contrary to what might be expected, the lexical density decreased. Interestingly, this is exactly what Vidakovic and Barker (2010) find in their analysis of written work at CEFR levels A1–C1 Skills of Life candidates’ written responses. Table 2 provides an overview of the scores.

Table 2. Summary of the means and standard deviations

	EFL: T1 N=31		EFL: T2 N=31		EFL: T3 N=31	
	Mean	s	Mean	s	Mean	s
Speech Rate	2.18	.33	2.43	.55	2.33	.30
Mean Word Length	3.68	.18	3.69	.15	3.80	.14
Type/Token Ratio	70.83	4.09	71.92	3.84	72.81	3.02
Lexical Density	39.90	4.60	38.65	4.09	38.34	3.85

Speech rate

The speech rate (words/second) ranged at T1 from 1.31 to 2.82, with a mean of 2.18; at T2 from 1.8 to 4.27, with a mean of 2.43; and at T3 from 1.78 to 2.91, with a mean of 2.33. So the speech rate went up considerably from T1 to T2 and then went down a little at T3. However, paired sample t-tests show that the rise between T1 and T2 is not statistically significant, nor is the drop between T2 and T3. However, the difference between the first and last recording is significant (t(30) = 2.73; p = .01).

Mean word length

The mean word length in characters ranged at T1 from 3.31 to 4.15, with a mean of 3.68; at T2 from 3.41 to 4.12, with a mean of 3.69; and at T3 from 3.46 to 4.1, with a mean of 3.80. So over time, the word length went up; the increase between T1 and T2 is not statistically significant, but the increase between T2 and T3 is (t(30) = 3.43, p < .01), and the overall increase over time is also significant (t(30) = 3.25, p < .01).

Type/token ratio

The type/token ratio ranged at T1 from 63 to 75.2, with a mean of 70.83; at T2 from 64 to 79.4, with a mean of 71.92; and at T3 from 68 to 78, with a mean of 72.81. Again we see a steady increase from T1 to T3, and as for the mean word length the difference between T1 and T2 was not statistically significant, nor was the increase between T2 and T3, but the overall increase is significant (t(30) = 2.99, p < .01).

Lexical density

The lexical density (percentage lexical words) ranged at T1 from 29 to 48, with a mean of 39.9; at T2 from 32 to 47, with a mean of 38.65; and at T3 from 29 to 45, with a mean of 38.34. Overall the lexical density is rather low, and it decreased steadily, but not statistically significantly, over time.

Lexical sophistication

Table 3 shows the lexical profiles of our students in the three recordings (percentages types and tokens in word lists 1, 2 and 3 and not in any lists). Given the nature of the prompts, it is hardly surprising that the figures for WL3 (academic words) remain low over the years. What is perhaps more surprising is that there does not appear to be shift from WL1 to WL2, suggesting that the learners keep operating within a relatively safe basic lexical environment.

Table 3. Lexical Sophistication measured with General Service List first (WL1) and second 1,000 (WL2) and Academic Word List (WL3) (in percentages)

	T1		T2		T3	
	Token	Type	Token	Type	Token	Type
WL1	82.92	62.25	85.35	54.89	84.19	58.09
WL2	4.16	10.24	3.87	13.00	4.18	10.82
WL3	1.38	3.93	1.97	6.23	1.69	6.81
Not in list	11.54	23.58	8.81	25.87	9.94	24.28

4. Discussion and conclusion

This project was carried out in order to find out whether it is possible to measure our students' spoken EFL development on the basis of these fairly crude measures. Statistics we used are within subject, i.e. we measure individual progress (over a whole group). While there are unmistakably very clear indications of development, even progress, that can be observed in the speech rate, the mean word length, and the type/token ratio, no conclusive evidence is provided by lexical density or lexical sophistication. Increased speech rate can be interpreted in terms of increased self-confidence, and ease of access to L2 vocabulary. Increased type/token ratio points to a greater lexical variation (cf. Johansson's (2008) lexical diversity), and an increased mean word length suggests a use of more sophisticated words.

However, these conclusions are not confirmed by the other two measures we used. There are two explanations for this. First, the nature of the tasks that we used for this study has obviously not stimulated the participants to explore (or use) their full academic lexical potential, as they were not required to address any specifically academic topics (unlike what they are required to do in the majority of their writing tasks, cf. de Haan and van der Haagen 2013a).

More importantly, it is not necessarily the case that the lexical progress of these advanced Dutch EFL learners becomes evident through an increase in lexical density, since their use of lexical words does not increase but even decreases. At first, this may seem rather surprising, because a high lexical density is usually considered to be characteristic of "the academic register" (i.e. advanced speech), as "it is expected that in spoken interaction in [formal] educational settings, lexical words are used to refer to entities or to situations, where in informal interactions these might be referred to by (deictic) pronouns or other function words" (Henrichs and Schoonen 2009: 4). So, speakers in academic settings are supposed to need many lexical words in order to achieve the clarity and explicitness which is required in the academic discourse. However, there is actually a logical explanation for the ascertained decrease in LD: the more advanced students become, the more complex sentences they form, and the more function words they have to use. Laufer and Nation (1995) also perceive this, though here with regard to written instead of spoken EFL production, saying:

> "[W]e could argue that the Lexical Density index does not necessarily measure lexis, since it depends on the syntactic and cohesive properties of the composition. Fewer function words in a composition may reflect more subordinate clauses, participial phrases and ellipsis, all of which are not lexical but structural characteristics of a composition. As the LD measure is influenced by the number of function words, this affects its validity." (309)

The upshot of this is that lexical density is not a valid method for measuring lexical progress. What is more, the decrease in lexical density that our students display suggests that it could actually be an indirect measure of syntactic advancement. The relationship between lexical density and syntactic complexity in spoken registers has also been commented on by Halliday (1989, 1994) and Biber (2006). Furthermore, Martin (1992) has shown that speakers in monologic expositions tend to hold the floor by using syntactically complex clauses that contain relatively few lexical items. This may be related to what hearers can process.

Given our findings, and comparing our results with what has been found in earlier studies on written EFL production, there is something to be said for comparing oral production with written production after all. We are tempted to conclude that EFL users go through comparable development stages at different levels in their oral production than in their written production. What is observed in written production at lower levels is observed for our students' oral production at higher levels. Our intention is therefore to combine lexical and syntactic features in a follow-up study, in order to measure EFL development in our students' spoken performance. One of our next steps is to examine how planned and unplanned production, task complexity (cf. Kuiken and Vedder 2007; Michel, Kuiken and Vedder 2007), and oral and written production compare.

Notes

[1] www.erk.nl

[2] The *Teacher's guide to the Common European Framework* is available online at http://www.pearsonlongman.com/ae/cef/cefguide.pdf.

References

Biber, D. (2006), *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: Benjamins.

Crookes, G. (1989), 'Planning and interlanguage variation', *Studies in Second Language Acquisition,* 11: 367-383.

de Haan, P. and M. van der Haagen (2012), 'Modification of adjectives in very advanced Dutch EFL writing: A development study', *The European Journal of Applied Linguistics and TEFL,* 1: 129-142.

de Haan, P. and M. van der Haagen (2013a), 'Assessing the use of sophisticated EFL writing: a longitudinal study', *Dutch Journal of Applied Linguistics,* 2: 16-27.

de Haan, P. and M. van der Haagen (2013b), 'The search for sophisticated language in advanced EFL writing: a longitudinal study', in: S. Granger, G. Gilquin and F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead*. Louvain-la-Neuve: Presses universitaires de Louvain. 103-116.

de Haan, P. and K. van Esch (2005), 'The development of writing in English and Spanish as foreign languages', *Assessing Writing,* 10*:* 100-116.

de Haan, P. and K. van Esch (2008), 'Measuring and assessing the development of foreign language writing competence', *Porta Linguarum,* 9: 7-21.

Halliday, M. A. K. (1989), *Spoken and Written Language*. 2nd ed. Oxford: OUP.

Halliday, M. A. K. (1994), *An Introduction to Functional Grammar*. 2nd ed. London: Edward Arnold.

Henrichs, L. and R. Schoonen (2009), 'Lexical features of parental academic language input: The effect on vocabulary growth in monolingual Dutch children', in: B. Richards, M. Daller, D. Malvern, P. Meara, J. Milton and J. Treffers-Daller (eds) *Vocabulary Studies in First and Second Language Acquisition: The Interface between Theory and Application*. Houndmills: Palgrave Macmillan. 1-22.

Hincks, R. (2010), 'Speaking rate and information content in English lingua franca oral presentations', *English for Specific Purposes,* 29: 4-18.

Johansson, V. (2008), 'Lexical diversity and lexical density in speech and writing: a developmental perspective', *Lund Working Papers in Linguistics,* 53: 61-79.

Kuiken, F. and I. Vedder (2007), 'Task complexity and measures of linguistic performance in L2 writing', *IRAL,* 45: 261-284.

Laufer, B. and P. Nation (1995), 'Vocabulary size and use: lexical richness in L2 written production', *Applied Linguistics,* 16: 307-322.

Martin, J. (1992), *English Text: System and Structure*. Amsterdam: Benjamins.

Michel, M., F. Kuiken and I. Vedder. (2007), 'The influence of complexity in monologic versus dialogic tasks in Dutch L2', *IRAL*, 45: 241-259.

Trofimovich, P. and W. Baker (2006), 'Learning second language suprasegmentals: effect of L2 experience on prosody and fluency characteristics of L2 speech', *Studies in Second Language Acquisition,* 28: 1-30.

Verheijen, L., P. de Haan and B. Los (2013), 'Information structure: the final hurdle? (The development of syntactic structures in (very) advanced Dutch EFL writing)', *Dutch Journal of Applied Linguistics,* 2: 92-107.

Vidakovic, I. and F. Barker (2009), 'Lexical development across second language proficiency levels: a corpus-informed study', paper presented at the BAAL Annual Conference 2009, Newcastle.

Vidakovic, I. and F. Barker (2010), 'Use of words and multi-word units in Skills for Life Writing examinations', *Cambridge ESOL: Research Notes,* 41: 7-14.

A longitudinal study of the syntactic development of very advanced Dutch EFL writing

Pieter de Haan and †Monique van der Haagen

Radboud University Nijmegen

Abstract

The study of EFL writing has so far not been able to go beyond the observation and discussion of group characteristics. While it has been possible to study developmental data, in the sense that writing products of students at various levels of proficiency were available for research, longitudinal EFL data were virtually non-existent. The LONGDALE project seeks to find an answer to the question how EFL writing develops over time. This article reports on a quantitative and a qualitative study of Dutch EFL writing, based on a modest amount of longitudinal data. Its aim is to find out if and how non-native writing develops over time, whether it develops in the direction of native writing, and whether individual students display individual developmental patterns. The answer to all three questions appears to be affirmative.

1. Introduction

Twenty years ago Tony Silva (1993) published an article reviewing the study of non-native writing. His conclusion was that non-native writing is on the whole less sophisticated than L1 writing, in that it is less fluent (i.e. shorter), less accurate (it contains more morphology and syntax errors), less effective (it receives lower holistic scores) and distinctly simpler in structure than native writing. Most of the articles reviewed by Silva reported on studies of ESL (English as a Second Language) data. Twenty years of study of written EFL (English as a Foreign Language) data have led to comparable conclusions for EFL writing (de Haan 1999; Granger 1998; Lorenz 1999). What is still not sufficiently known is how non-native writing develops over time. The new LONGDALE project (Granger 2009) seeks to provide an answer to this question.

Finding an answer to this question is especially relevant in the Dutch university context. Most university programmes in the Netherlands do not, as a rule, provide English language courses to the students, on the assumption that they have acquired a sufficient amount of English to cope with English study materials. Dutch secondary school leaving exam requirements the level B2 (vantage) in the CEFR (Melissen 2007). Students of English, by contrast, are expected to develop their command of English to CEFR level C2 (mastery) by the time they graduate. They are the ones that will become what we like to call EFL professionals. We use this term to describe non-native speakers of English who are employed as language teachers, language trainers, translators, editors, usually in a non-native English environment.

There are a number of linguistic observations that are relevant in the study of writing development in general, and the development of non-native writing in particular. The study of EFL writing in the past has shown it to be characterised by the occurrence of a large number of speech elements (de Haan and van der Haagen 2013). The occurrence of speech elements in written language has also been shown to be a feature of novice native writing (Shaughnessy 1977). It is as though formal academic writing is a foreign language, even to native speakers (Hamp-Lyons 2011).

Earlier studies have shown that speech tends to be characterised generally by a more verbal style, and an abundant use of personal pronouns, while formal writing has a definitely more nominal style (de Haan 2001, 2002). We will therefore assume that for writing (whether native or non-native) to be judged "sophisticated", it should be characterised by a nominal, rather than a verbal style. In other words, sophistication is associated with writing, because writing reflects learning, i.e. more educated language use. At the same time, there is increasing insight in the nature of cohorts of learners as dynamic systems, even at advanced stages, implying that group developmental dynamics are an amalgamation of the individual group members' development (de Bot, 2008; de Bot, Lowie, and Verspoor, 2007; Dörnyei, 2007).

What this study aims to address is the question whether we can observe an increase of the use of sophisticated language by non-native writers of English, as it is reflected in the use of specific word classes and word class combinations. More specifically, we want to address the following research questions:

1. Are there any salient differences between native and EFL writing in the occurrence of nouns, verbs, and pronouns?
2. What does the occurrence of these word classes in EFL writing tell us about the level of sophistication?
3. Do the group findings for the EFL writers reflect individual EFL writing?

2. Methodology

For the current study we examined the data collected from a single cohort of students of English at Radboud University Nijmegen. These students started their studies in September 2011. Students take writing courses in each of the three BA years. In these writing classes they are specifically taught such things as the formulation of thesis statements, the construction of paragraphs, avoiding logical fallacies, avoiding possible plagiarism. In particular the first year writing course is devoted to practicing the use of cohesive devices and logical connectors. No explicit attention is paid to the syntactic make-up of the sentences in their written work, other than the correct positioning of adverbials, although students do practice sentence embedding and the use of relative clauses. In the second year and the third year courses gradually less attention is paid to these mechanics, and attention is gradually shifted to tackling more complex and demanding topics.

We collected six written assignments over a period of five months, starting with a personal paper in September, during their first writing class. We collected three in-class assignments, and three home assignments, according to the scheme presented in Table 1. The EFL material was compared to a selection of essays from the LOCNESS corpus, which served as native reference material. The LOCNESS material was composed of UK and US argumentative student essays on a variety of topics, both literary and non-literary, produced in timed and untimed conditions. The EFL students' in-class assignments were completed in thirty minutes.

The issue of EFL data collection has received some attention in the past. De Haan and van Esch (2004, 2005, 2007, 2008) collected longitudinal data of a number of cohorts of Dutch students of English and Dutch students of Spanish, in an attempt to chart basic quantitative foreign language writing development features. They found clear signs of development in the Spanish students' writing, in terms of word length, sentence length, and type/token ratio scores. For the English students' writing no such development was witnessed, which was attributed to the fact that these students were very advanced EFL users, as opposed to the Spanish students, who had not had any training in Spanish prior to university. The advanced students of English did not display any development in these basic features.

Earlier, it had been suggested (Ortega, 2003; Shaw and Liu, 1998) that collecting longitudinal data at relatively short intervals (i.e. shorter than nine months) might not be useful for the study of syntactic development, as students seldom displayed more advanced syntactic behaviour at such intervals. However, Hayes, Hatch and Silk (2000) argued that students are unlikely to show their full potential in any arbitrarily selected writing assignment, and therefore the aim should be to collect as many written data as possible. It is with the latter in mind that we decided to collect so many writing assignments from this cohort of students. Following another suggestion by Hayes, Hatch and Silk, we also decided to collect writing assignments that covered a wide variety of tasks and topics, in order to ascertain earlier findings (de Haan and van der Haagen 2012, 2013) that the writing task and the topic tend to influence the students' lexical and syntactic behaviour.

Table 1 shows the data that was used for this study. We collected six texts from our first year students within a period of five months. We collected a comparable amount of native students' data, which we used for reference. It should be noted that three of the writing assignments were completed in class; these were 30-minute assignments that the students completed without access to reference books. The other three assignments were home assignments; there were no time limits for these, which is reflected in the mean essay lengths. The general mean essay length for the non-native writers in this study is 467 words. The mean essay lengths for the various writing tasks reflect the timed conditions in texts 1, 2, and 5. The topics for the writing assignments were as follows:

Text 1	“My expectations of the coming year”
Text 2	“1st year students can take specific steps to become successful in college”
Text 3	British literature – analysis of a short story
Text 4	British literature – analysis of student’s own sonnet
Text 5	“The Dutch government should spend more money on primary / less money on tertiary education”
Text 6	Take home exam American literature – American Romanticism

Table 1. Data collection and basic statistics

	text 1	text 2	text 3	text 4	text 5	text 6	native reference corpus
	Sept 2011	Sept 2011	Oct 2011	Dec 2011	Dec 2011	Jan 2012	LOC-NESS
	in class	in class	at home	at home	in class	at home	UK / US
# essays	96	94	86	82	66	78	58
# words	27,911	36,526	47,114	45,880	19,477	57,347	47,495
mean # words/ essay	291	389	548	560	295	735	819

The data were automatically PoS tagged with the NLP Studio tagger (Northedge 2006), after which the frequency distribution of the tags was calculated. We did not specifically train the tagger, as we used both native and non-native material, which may have affected the tagger’s accuracy (for a discussion about the accuracy of PoS taggers, see e.g. Manning 2011). As we were particularly interested in finding out about individual students’ behaviour, we selected the work of a small number of non-native students who contributed all six assignments, and compared the findings for these students to the native speakers’ data.

3. Results

3.1 Quantitative analysis

Table 2 shows the relative frequency distribution of the most common PoS tags for the native essays and those written by four of the Dutch students. We decided to restrict the more detailed analysis to four students, as we wanted to subject their writings to a detailed qualitative analysis in a following stage (see Section 3.2). The figures are presented in decreasing order of frequency found in the native essays. The mean essay length for the four non-native students is 457 words per essay, which appears to reflect the general mean essay length of the entire group of non-native students (467 words). The table shows a significant non-native underuse of nouns (crosstabulation yields a standardised residual score of –6.2), and a significant non-native overuse of personal pronouns (standardised residual score 8.9). The underuse of nouns by our non-native writers points to a lesser level of sophisticated language use. This is exactly what we expected on the basis of de Haan's (2001) findings: non-native writing is characterised by a large number of speech features.

Table 2. Distribution of most common PoS tags in native essays and non-native essays written by four Dutch students (percentage scores in decreasing order of frequency in the native essays)

tag	native N=47850	non-native N=10968
Noun	27.5	24.2
Lexical verb	16.3	16.7
Preposition / subordinator	13.7	12.8
Determiner	11.6	13.1
Adjective	9.1	8.7
Adverb	5.4	5.5
to as prep. or infin. marker	3.0	2.9
Coordinator	3.0	3.3
Personal pronoun	2.7	4.2
Modal auxiliary	2.2	1.6
other	5.5	7.0

We were interested in finding out how individual students behave syntactically, and how they behave in the various assignments. Table 3 shows the PoS tag frequency distribution of these four individual students. They are identified by their LONGDALE identification numbers. Student RAD1104 did not submit all six assignments; therefore we included student RAD1105's texts.

Table 3 shows that there are large differences between these four individual students' scores. There are obviously considerable differences in total student output (ranging from 2417 words for RAD1103 to 3121 words for RAD1105), which might be taken as an indication of general proficiency differences. Also, we see individual differences in the distribution of the various word classes. In the use of nouns and personal pronouns, RAD1101 has almost exactly the same relative scores as the native students, while the other three non-

Table 3. Distribution of most common PoS tags in native essays and non-native essays written by four Dutch students (percentage scores in decreasing order of frequency in the native essays)

tag	native N= 47850	RAD1101 N= 2913	RAD1102 N= 2517	RAD1103 N= 2417	RAD1105 N= 3121
Noun	27.5	27.0	23.4	24.0	22.2
Lexical verb	16.3	14.4	17.2	16.5	18.8
Preposition / subordinator	13.7	13.8	12.8	13.0	11.6
Determiner	11.6	13.0	12.9	14.0	12.8
Adjective	9.1	9.8	8.5	8.9	7.6
Adverb	5.4	4.9	5.8	5.1	6.3
to as prep. or infin. marker	3.0	2.4	2.7	3.7	3.3
Coordinator	3.0	3.3	3.6	3.1	2.9
Personal pronoun	2.7	2.3	5.2	4.3	5.0
Modal auxiliary	2.2	1.8	1.3	0.9	2.1
other	5.5	7.3	6.6	6.6	7.4

native students underuse nouns, and overuse personal pronouns. In the use of lexical verbs, on the other hand, it is RAD1101 who underuses them, while the other three students overuse verbs to a greater or lesser extent.

Tag bigrams have been shown to be good indicators of basic syntactic patterns in speech and (formal) writing (de Haan 1999, 2001). It has been shown, for instance, that noun-noun combinations are more typically used in writing than in speech. So the occurrence of specific tag bigrams can be taken to represent the adequacy of EFL writing. Table 4 shows the most common tag bigrams.

Table 4 reveals clear differences between the native essays and the non-native essays. Even though these ten tag bigrams account for only slightly under half of all the combinations, we see, again, clear differences between the native and the non-native writers. There is a notable underuse of noun-noun combinations (standardised residual score –9.6) and preposition-noun combinations (standardised residual score –5.3) by the non-native writers, and an overuse of determiner-noun combinations (standardised residual score 4.2). However, when we break down our non-native writers' results for the various writing tasks, we can observe highly interesting differences between the first text (text 1), written in September, and the last text (text 6), written four months later, in January. These figures are presented in Table 5.

Table 4. Distribution of most common tag bigrams in native essays and non-native essays written by four Dutch students (percentage scores in decreasing order of frequency in the native essays)

tag	native N=47849	non-native N=10967
determiner – noun	7.2	8.5
noun – preposition	6.9	6.1
adjective – noun	6.1	5.6
preposition – determiner	4.9	5.3
preposition – noun	4.1	2.9
noun – noun	4.1	1.9
noun – verb	4.0	3.3
determiner – adjective	3.1	3.4
verb – preposition	3.0	2.5
verb – determiner	2.8	3.4
other	53.8	57.1

Table 5. Distribution of most common tag bigrams in native essays and early and more advanced non-native essays (percentage scores in decreasing order of frequency in the native essays)

tag	native N=47849	non-native t1 N=1129 Sept. 2011	non-native t6 N=2684 Jan. 2012
determiner – noun	7.2	6.1	9.2
noun – preposition	6.9	5.3	7.6
adjective – noun	6.1	3.9	5.7
preposition – determiner	4.9	4.6	5.7
preposition – noun	4.1	1.6	4.1
noun – noun	4.1	1.0	3.1
noun – verb	4.0	1.4	3.8
determiner – adjective	3.1	2.0	3.5
verb – preposition	3.0	2.6	2.2
verb – determiner	2.8	3.4	3.5
other	53.8	68.1	51.6

Table 5 shows how an initial underuse of determiner-noun in September is turned into an overuse in January. It is also shown that the underuse of preposition-noun and noun-noun combinations is even more dramatic in September, but appears to be moving towards the native reference scores in January. The slight underuse of noun-verb combinations in Table 4 is now revealed to display the same developmental pattern as the preposition-noun and noun-noun combinations. Clearly, an underuse of nouns affects all the combinations in which it is found.

Another interesting observation is of course the decrease in the number of "other" scores. In September the non-native writers use the ten most frequent tag bigram combinations in less than one third of their texts. By January this has increased to almost half, which we take to be an indication of a growing awareness of more appropriate combinations, which is undoubtedly due to the grammar and writing classes, but also to their academic course reading.

However, this still does not give us an adequate idea of individual students' syntactic behaviour in the various writing assignments. Ideally, the figures for the PoS scores and tag bigrams of all six texts for each of the four non-native students should be presented here, but due to space restrictions we have limited the detailed presentation and discussion to the PoS tags scores of two students, viz. RAD1101 and RAD1102. Table 6 presents the frequency scores of the tags for RAD1101 in each of the six texts. Table 7 presents the frequency scores for RAD1102 in each of the six texts.

Table 6. Distribution of most common PoS tags in native essays and in RAD1101 (percentage scores in decreasing order of frequency in the native essays)

tag	native N= 47850	text 1 N= 245	text 2 N= 369	text 3 N= 529	text 4 N= 608	text 5 N= 353	text 6 N= 809
Noun	27.5	19.6	21.1	31.4	24.0	26.1	31.8
Lexical verb	16.3	16.3	17.9	11.3	17.3	12.5	12.9
Preposition / subordinator	13.7	15.5	11.9	13.8	12.3	13.9	15.1
Determiner	11.6	11.4	11.9	14.7	13.0	8.5	14.7
Adjective	9.1	9.4	10.8	8.9	9.2	14.7	8.4
Adverb	5.4	6.9	7.6	3.6	5.6	6.5	2.6
to as prep. or infin. marker	3.0	2.9	4.3	0.8	2.3	3.4	2.2
Coordinator	3.0	4.1	3.5	4.0	3.6	2.3	2.8
Personal pronoun	2.7	6.9	3.0	1.3	1.8	2.3	1.6
Modal auxiliary	2.2	0.8	2.7	1.3	2.0	4.5	0.7
other	5.5	6.2	5.3	8.9	8.9	5.3	7.2

Table 7. Distribution of most common PoS tags in native essays and in RAD1102 (percentage scores in decreasing order of frequency in the native essays)

tag	native N= 47850	text 1 N= 246	text 2 N= 316	text 3 N= 540	text 4 N= 435	text 5 N= 246	text 6 N= 734
Noun	27.5	18.7	18.4	24.1	24.6	25.2	25.5
Lexical verb	16.3	21.5	22.2	18.3	13.1	18.7	15.9
Preposition / subordinator	13.7	10.6	10.4	13.7	15.6	11.0	12.8
Determiner	11.6	8.9	7.3	18.3	14.5	9.3	12.9
Adjective	9.1	8.1	7.6	10.6	5.5	10.6	8.4
Adverb	5.4	9.3	9.2	4.4	4.6	6.1	4.6
to as prep. or infin. marker	3.0	2.8	3.8	1.9	1.8	3.3	3.1
Coordinator	3.0	6.1	2.5	3.1	3.4	3.3	3.8
Personal pronoun	2.7	9.3	8.5	3.1	6.2	4.1	3.7
Modal auxiliary	2.2	2.0	3.2	0.9	0.7	2.4	0.5
other	5.5	2.7	6.9	1.6	10.0	6.0	8.8

The frequency data for nouns and personal pronouns are interesting to look at, as we saw earlier that nouns tended to be underused by the non-native writers, and pronouns tended to be overused. The data in Tables 6 and 7 show how RAD1101's use of nouns in quantitative terms hovers between underuse and overuse. The only text in which the score resembles the native score is text 5. RAD1102, on the other hand, shows a steady increase in the use of nouns from text 3 onward. When we look at the use of personal pronouns, we see different patterns: neither student shows a steady decrease in the use of pronouns, but RAD1101's use of personal pronouns is inversely proportional to RAD1102's in texts 2 and 4. We will come back to this in Section 3.2.

3.2 Qualitative analysis

We can now assess the data presented in the previous subsection in the light of our research questions. We find that our non-native writers show an overall underuse of nouns, and an overuse of personal pronouns. This would seem to confirm earlier general findings that non-native written English tends to have more speech characteristics, one of them being a less frequent use of nouns, and nominal constructions. The latter is confirmed by our examination of the occurrence of tag bigrams featuring nouns. These, too, show a distinct underuse by the non-native writers. However, comparing early and later non-native writing performance reveals that underuse in most cases is caused by dramatic underuse by the beginning writers, but that the figures for the same writers a few month later are much more like those of the native writers. This might lead us to the conclusion that the students whose data we analysed shows signs of development in the direction of native usage. What is especially telling in this respect is the observation that in case of tag bigrams the number of "other" categories decreased dramatically in a five-month interval, indicating a growing awareness in the non-native writers of more common, appropriate, combinations.

Still, the data were analysed in more detail, and this more detailed analysis revealed interesting individual differences between two students. These differences need to be commented on. We noted individual differences in the use of the nouns and the personal pronouns. If we start with the latter, it should be borne in mind that the writing task definitely played a role here, particularly in texts t1 and t2. Both students displayed a massive overuse of personal pronouns in t1, but only RAD1102 overused personal pronouns in t2. The topic in the first text was the students' expectation of the coming year, where it can be expected that they would refer to themselves frequently. This turns out to be the case, witness the following extracts from both these students' first essays.

> ... I basically start from scratch. First, I need to explore Nijmegen, since it is a city that I had never visited before the introduction week. Besides that, I have to adjust myself to a whole new way of living, because I am now entering the world of small rooms and living on your own. ... (RAD1101, t1)

> I expect the coming year to be very busy and fast-paced, because of all the work that will need to be done. I expect time before a deadline will go way too fast and I will hae to try to run after in and get everything done. Luckily, I find the courses very interesting so far, so that makes it a lot easier. I do not think studying is all about studying, though. ... (RAD1102, t1)

However, the prompt for the second writing task (t2) invited students to list specific steps students can take to ensure they are successful in college. Many students took this as an invitation to address the imagined reader directly, giving him all kinds of advice. RAD1102 was one of those:

> Every student wants to be successful at university. Studying is expensive, so everyone wants to make the best of it. Fortunately, there are certain steps you can take to make your time at university as successful as possible.
>
> First of all, work hard. It is said very often, but it is true. You cannot pass your exams if you do not do anything. Keeping up with your assignments is half the work, so make sure you know what you

> have to do every week and do your assignments with care, so you will actually learn something. ... (RAD1102, t2)

In contrast, RAD1101 chose not to address the reader directly, thus creating a more academically appropriate detached tone:

> ... First, a student must be willing to show up at every class, let it be either a lecture or a seminar. Showing up for class really helps a student understand the subject for tuition, and therefore the subject for homework. Besides, showing up for class shows a teacher that one is truly interested in the study and willing to get the most out of it.
>
> Second, setting up a clear schedule is key to getting a good view of homework and the study as a whole. Do not have the intention to do certain homework the day before it needs to be done. Instead, do it a couple of days in advance so there is enough time for other projects or outside interests. ... (RAD1101, t2)

Interestingly, the word *you* occurred eighteen times in RAD1102, t2, and *your* occurred a further twelve times. Neither word occurred at all in RAD1101, t2. Apparently RAD1101 is more aware of what might be called the appropriate academic register than RAD1102. It is especially these issues that the first year writing classes address, so we would expect students to display gradually more signs of formal academic writing.

One other specific observation that we would like to comment on is the distribution of nouns and lexical verbs. With one exception (t4), RAD1101 uses relatively more nouns than RAD1102, but fewer verbs. Relating the figures to the native reference figures we see a more erratic pattern in RAD1101 than in RAD1102; the latter consistently uses fewer nouns. Considering our earlier observation, that an underuse of nouns was likely to point to a more speech-like production, this suggests that RAD1102 is not so successful in acquiring the right academic style, as RAD1101. This seems to be confirmed by what they produce in t5, where they have to argue a non-trivial point. Even though there is hardly any difference between the two in the use of nouns in numerical terms, there are great differences in the way the points are presented, as the two excerpts below show. It should be borne in mind that this was an in-class, timed assignment. Below are the two opening paragraphs of their short essays.

> There are several reasons why primary education is more important than tertiary education and thus the Dutch government should spend less money on tertiary education and more money on primary education. The reason why primary education is the most important type of education, is because it should give children a fundamental amount of knowledge for the rest of their lives; it should provide children with a fluent transition to secondary school; and it should create equal chances for everyone.

> First, primary education should be used to give children an ideal basis for the rest of their lives, even when there are differences in capabilities children have. Whereas tertiary education is only destined for a smaller group of people, primary education is for everyone. As a result, people who are not smart enough to go to university should be able to get a solid basis of knowledge for the rest of their lives at primary school. ... (RAD1101, t5)

> Primary education is the basis of education. The things we learn in primary school are important to make a living. If the primary school system fails, children are already behind on knowledge. If the Dutch government wants to maintain a high standard of education, they should start right at the base, with primary education.
>
> In primary school, children do not just learn to read or write, they learn all kinds of other things that are important for their future. They learn how to work together, communicate and plan their time, and learn important values that are useful for the rest of their lives. Primary schools are also a place where children learn about their culture, they hear stories and songs that belong to their culture. ... (RAD1102, t5)

RAD1101 uses a proper introduction to the topic, and prepares the reader for what will be elaborated on in the remainder of the essay (which is one of the issues dealt with in the writing class). RAD1102, on the other hand, does not address the topic properly, and displays a number of speech elements, like short sentences. We are possibly looking at a relation between immature language use and immature writing skill. Clearly, a qualitative analysis of the students' texts gives a better insight into language use and writing skill, but at this stage is hard to establish whether RAD1102's poor opening paragraph is due to a poor command of English, or whether the student's unsophisticated use of English is due to the struggle to construct a proper opening paragraph.

4. Conclusion

We conclude from this study that our three research questions can be answered, in the sense that we find that non-native writing is different from native writing as far as the use of certain word classes goes. More particularly, the overuse of pronouns and the underuse of nouns indicates a less sophisticated written production by the non-native writers overall as these are characteristic of speech, rather than (formal) writing. Finally, there are clear differences between individual EFL writers in the development of their written production, which appear to indicate the level of success in the acquisition of an appropriate formal academic writing style.

However, there remain a number of problems. We noticed in the introduction that we had deliberately included a number of very different writing products in our study, in order to get a broad picture of what our students are capable of. This has the drawback that at least some of our findings are affected by the writing task, for which it is difficult to control the data. More advanced writing cannot be directly related to progress in time, but will also be affected by the demand made on the student to formulate their thought and arguments. Also, the amount of course reading that feeds into their production, for instance in the case of take-home exam papers, will contribute to the level of sophistication in the students' writing.

Furthermore, we have found that the automatic tagger produced errors at certain points. It would be naive to assume that the tagger will produce the same amount of errors in all the texts, and that therefore these errors can be ignored. We were surprised to find, for instance, that the combination *American Romanticism*, which occurred a few times in the non-native texts (t6) was tagged as noun–noun. Therefore, a careful check of the tags assigned would be in order.

Finally, as the discussion section has shown, we will have to look at more individual students' essays qualitatively. While we believe it is fair to say that both quantitative syntactic analysis and qualitative analysis reveal the level of sophistication in EFL writing, we are convinced that further qualitative analysis of individual students' successive writing products will provide us with a better view of what non-native writing development at an advanced level amounts to.

References

de Bot, K. (2008), 'Second language development as a dynamic process', *Modern Language Journal*, 92: 166-178.

de Bot, K., W. Lowie and M. Verspoor (2007), 'A Dynamic Systems Theory approach to second language acquisition', *Bilingualism: Language and Cognition*, 10: 7-21.

de Haan, P. (1999), 'English writing by Dutch-speaking students', in: H. Hasselgård and S. Oksefjell (eds) *Out of Corpus: Studies in Honour of Stig Johansson*. Amsterdam: Rodopi. 203-212.

de Haan, P. (2001), 'Aspects of the syntax of spoken English', in: K. Aijmer (ed.) *A Wealth of English: Studies in Honour of Göran Kjellmer*. Gothenburg: Gothenburg University Press. 47-56.

de Haan, P. (2002), 'The non-nominal character of spoken English', in: L. E. Breivik and A. Hasselgren (eds) *From the COLT's Mouth ... and Others'. Language Corpora Studies in Honour of Anna-Brita Stenström*. Amsterdam: Rodopi. 59-69.

de Haan, P. and M. van der Haagen (2012), 'Modification of adjectives in very advanced Dutch EFL writing: a development study', *The European Journal of Applied Linguistics and TEFL*, 1: 129-142.

de Haan, P. and M. van der Haagen (2013), 'Assessing the use of sophisticated EFL writing: a longitudinal study', *Dutch Journal of Applied Linguistics,* 2: 16-27.

de Haan, P. and K. van Esch. (2004), 'Towards an instrument for the assessment of the development of writing skills', in: U. Connor and T. Upton (eds) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi. 267-279.

de Haan, P. and K. van Esch (2005), 'The development of writing in English and Spanish as foreign languages', *Assessing Writing*, 10: 100-116.

de Haan, P. and K. van Esch (2007), 'Assessing the development of foreign language writing skills: syntactic and lexical features', in: E. Fitzpatrick (ed.) *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodopi. 185-202

de Haan, P. and K. van Esch (2008), 'Measuring and assessing the development of foreign language writing competence,' *Porta Linguarum*, 9: 7-21.

Dörnyei, Z. (2007), *Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies*. Oxford: OUP.

Granger, S. (2009), LONGDALE. Retrieved 22 October, 2010, from http://www.uclouvain.be/en-cecl-longdale.html.

Granger, S. (ed.) (1998). *Learner English on Computer*. London: Addison Wesley Longman.

Hamp-Lyons, L. (2011), 'Assessing the ineffable', paper presented at the Writing Assessment in Higher Education Symposium, Amsterdam.

Hayes, J. R., J. A. Hatch and C. M. Silk (2000), 'Does holistic assessment predict writing performance? Estimating the consistency of student performance on holistically scored writing assignments', *Written Communication*, 17: 3-26.

Lorenz, G. (1999), *Adjective Intensification – Learners versus Native Speakers*. Amsterdam: Rodopi.

Manning, C. D. (2011), 'Part-of-Speech tagging from 97% to 100%: Is it time for some linguistics?', in: A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing* (Vol. 1). Berlin Heidelberg: Springer. 171-189.

Melissen, M. (2007), *Exameneisen havo-vwo nieuwe stijl 2007.* Alphen aan den Rijn: Kluwer.

Northedge, R. (2006), *Statistical parsing of English sentences*. Retrieved from http://www.codeproject.com/Articles/12109/Statistical-parsing-of-English-sentences.

Ortega, L. (2003), 'Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college–level L2 writing', *Applied Linguistics*, 24: 492-518.

Shaughnessy, M. P. (1977), *Errors and Expectations: A Guide for the Teacher of Basic Writing*. New York: Oxford University Press.

Shaw, P. and E. Liu (1998), 'What develops in the development of second-language writing?', *Applied Linguistics*, 19: 225-254.

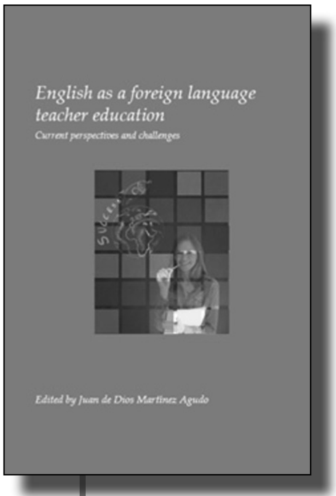
Silva, T. (1993), ‘Toward an understanding of the distinct nature of L2 writing: The ESL research and its implication’, *TESOL Quarterly*, 27: 657-677.

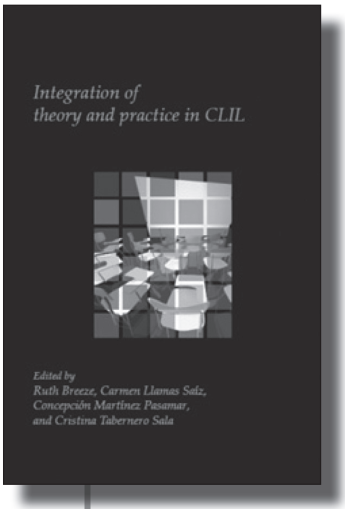
STUDIES IN SLAVIC AND GENERAL LINGUISTICS

INNOVATION IN TRADITION

TÖNNIES FONNE'S RUSSIAN-GERMAN PHRASEBOOK (PSKOV, 1607)

PEPIJN HENDRIKS


English as a foreign language teacher education
Current perspectives and challenges
Edited by Juan de Dios Martínez Agudo



STUDIES IN SLAVIC AND GENERAL LINGUISTICS
EDITED BY
EGBERT FORTUIN
PETER HOUTZAGERS
JANNEKE KALSBEEK
SIMEON DEKKER
DUTCH CONTRIBUTIONS TO THE FIFTEENTH INTERNATIONAL CONGRESS OF SLAVISTS
MINSK
AUGUST 20-27, 2013
LINGUISTICS