The Corpus Linguistics Discourse

In honour of Wolfgang Teubert

EDITED BY
Anna Čermáková
Michaela Mahlberg

Studies in Corpus Linguistics 87

JOHN BENJAMINS PUBLISHING COMPANY

The Corpus Linguistics Discourse

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/scl

Volume 87

The Corpus Linguistics Discourse. In honour of Wolfgang Teubert
Edited by Anna Čermáková and Michaela Mahlberg

The Corpus Linguistics Discourse

In honour of Wolfgang Teubert

Edited by

Anna Čermáková
Michaela Mahlberg
University of Birmingham

John Benjamins Publishing Company
Amsterdam / Philadelphia

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.87

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2018036404 (PRINT) / 2018050422 (E-BOOK)

ISBN 978 90 272 0175 1 (HB)
ISBN 978 90 272 6326 1 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Introduction

Anna Čermáková and Michaela Mahlberg
University of Birmingham

> Probably no society has ever been more concerned with meaning than the one in which we live. Never before have so many people felt such an urge to make sense of the world they live in and of the lives they are leading. They find this sense not so much in themselves as in the discourse […] (Teubert 2010: 1)

The aim of this book is to make a reflective contribution to the corpus linguistics discourse. While any publication, talk or tweet that says something about corpus linguistics makes a contribution to this discourse, the present book aims to explicitly reflect on some of the crucial concepts and approaches that affect how we negotiate meanings in and push the boundaries of our discipline. Since the beginning of the 'corpus revolution' triggered by advances in computing and initially most strikingly visible in innovations in lexicography, corpus linguistic methods, tools, resources and applications have developed to cover a wide variety of contexts. Still, it seems relatively little time is spent reflecting critically on the underlying assumptions that our field is based on. With this book we want to make a contribution to this reflection. Although 'discourse' appears in the title, this is not a book about corpus-assisted approaches to discourse analysis. The book is about the discourse in corpus linguistics – that is we use a corpus linguistic conceptualisation of the notion of discourse to apply it to our own field.

The book is a publication in honour of Wolfgang Teubert. Wolfgang made a significant contribution to corpus linguistics through the way in which he has approached the concept of meaning and its intricate relationship with the notion of discourse. He states: "[c]orpus linguistics studies languages on the basis of discourse" (Teubert 2004: 100). The term 'discourse' is widely used and can be employed with sufficiently vague meaning as needed. However, a critical engagement with the concept is fundamental to corpus linguistics. A basic definition by Brown and Yule (1983: 1) that describes discourse analysis as the "analysis of language in use" helps to emphasise a crucial point: corpora are electronic collections of naturally occurring textual data. In other words – corpora provide "*evidence* of language in use" (Tognini-Bonelli 2001: 47, our emphasis). Teubert refers to

https://doi.org/10.1075/scl.87.01cer

language in use as 'the discourse', that is "the entirety of everything that has been said and written by the members of the discourse community" (Teubert 2010: 1). In his approach, he emphasises how reality is constructed in the discourse through the process of exchanging and sharing verbal interactions among people, that is the members of a discourse community. Teubert describes the discourse as a "limitless, all-encompassing blog uniting humankind" (Teubert 2010: 2). People, as members of society, incessantly contribute to this blog. As Teubert (2010: 3) points out: "[w]hen we talk, we never start at point zero. We react to things that have been said before". This approach to discourse has direct implications for the definition of meaning in discourse. For Teubert (2010: 188), "[t]he meaning of a lexical item is everything that has been said about it in the discourse under investigation".

A holistic approach to the discourse as the entirety of language in use addresses corpus linguistic concerns in an abstract and theoretical manner. In practical terms, the evidence that is examined by corpus linguistic studies consists of samples of language in use. A large amount of work in corpus linguistics deals with the collection of data on the one hand and on the other hand, with the development of software and tools to process the data, as well as the linguistic analysis of it. Some corpus linguists will engage in all three types of activity. Often, however, the field will develop through collaboration and division of labour. A challenge that comes with this division of labour is to ensure a productive discourse between members of the corpus linguistic community where key meanings are shared. Teubert's approach to meaning puts particular emphasis on the notion of 'paraphrases' (e.g. Teubert 2010: 204–205). To paraphrase meanings is a way of telling others what a word means. Such paraphrases show how meaning is not fixed but negotiated by the discourse community. Because of the division of labour in corpus linguistics, key meanings for the discipline need to be defined, paraphrased and negotiated. Teubert (2005) illustrates this process in practice, in his position paper pointedly entitled "My version of corpus linguistics".

That not all corpus linguists share the same meanings to define the field became particularly apparent in the debate around 'corpus-based' versus 'corpus-driven' approaches that was often aired at corpus linguistic conferences in the early years of the 21st century. Tognini-Bonelli (2001) uses the terms to distinguish corpus approaches that begin from a specific theoretical position from approaches that set out to derive theoretical insights from the corpus data. While not necessarily expressed in the same terminology, the distinction is also reflected in approaches that prefer to see corpus linguistics as a methodology versus those that are interested in the theoretical arguments of corpus linguistics. A snapshot of the diverse opinions on the status of corpus linguistics is captured in a special issue of the *International Journal of Corpus Linguistics*. As guest editor for the journal, Worlock Pope (2010) turned a discussion that originated on the Corpora List into

a collection of articles debating the kind of methodological or theoretical contributions corpus linguistics is able to make. Less in the form of a debate, but still giving space to individual positions on the state of the field, Viana, Zyngier and Barnbrook (2011) compiled a collection of interviews with prominent researchers in corpus linguistics to collect their views on fundamental questions in the field.

In the present book, we approach the corpus linguistics discourse through a rather specific sample of it. We have chosen contributions to address what we regard as two key notions in corpus linguistics: 'meaning' and 'methods'. Both meaning and methods are intricately connected with the concept of discourse and they are equally broad terms. Our selection of contributors gives them very specific meanings with crucial theoretical implications for the foundations of corpus linguistics.

Meaning

The main result of the corpus revolution in lexicography was an innovative approach to meaning. As Sinclair (1987) showed in an account of the Cobuild project, new corpus-informed dictionaries reflected the crucial relationship between meanings and frequencies. Equally, words were no longer seen as the main units of meaning, but the importance of the patterns in which words occur became recognised as fundamental to the description of meaning. In present-day corpus linguistics, this approach to meaning is seen in a variety of terms such as 'lexical bundles', 'clusters', 'concgrams', 'phrase frames', etc. In his contribution to the present volume, Stubbs states: "[a]rguably the most far-reaching discovery of computer-assisted corpus analysis is that it is not individual words, but recurrent collocations, which are the basic units of meaning." However, he sets out to show that this idea (like other central ideas in corpus linguistics) originated much earlier, while corpus methods made for a much speedier development of the conceptualisation of units of meaning. Similar to Stubbs, Krishnamurthy is also concerned with the history of corpus linguistics. His particular focus is the units of language and their relationship to meaning. Unlike Stubbs', his argument goes beyond the word and considers linguistic units from characters to morphemes through various levels of language till he arrives at the discourse level. Going back to ancient linguistic questions, Krishnamurthy thus discusses how corpus linguistics has redefined traditional linguistic units and defined new ones.

The magnitude that Teubert ascribes to *the* discourse as an all-encompassing phenomenon links it to the object of study in discourse analysis in the more traditional sense. The 2nd edition of the three-volume *The Handbook of Discourse Analysis* (Tannen, Hamilton & Schiffrin 2015) clearly shows the extensive nature

of the field. Discourse analysis is not only a linguistic discipline but examines the language in relation to "the individual, society, and culture", in other words it examines "the real-world contexts that are in part created by discourse as they are sites for its use" (Tannen 2015: xx). What unites most research in discourse analysis is the qualitative nature of the analysis (Baker & Ellece 2011: 32). The adoption of corpus methods can thus be seen as contributing to a "quantitative turn" (e.g Sampson 2013).

Partington (2010) uses the term 'Corpus-Assisted Discourse Study' (CADS) to refer to the analysis of discourse shaped by corpus linguistic methods. He describes the aim of CADS as "to uncover, in the discourse type under investigation, what we might call *non-obvious meaning* – that is, meaning which might not be readily available to perusal by the naked eye" (Partington 2010: 88, emphasis in the original). In the present volume, Partington and Zuccato illustrate the CADS approach by studying the UK's debate on membership of the European Union. They begin by revisiting a study by Teubert (2001) on EU-sceptic attitudes in the UK in 2000 and take their research to more recent debates up until immediately after the Referendum on Brexit in June 2016. Partington and Zuccato's analysis shows how different voices interact and compete to be heard.

Teubert (2010: 1) states that "[i]n principle, everyone has a voice in the discourse" however, in reality these voices are not distributed evenly across the discourse. The chapter by Philip in the present volume highlights what measures might be taken to affect how specific voices are represented. Philip's interest is in a Style Guide proposed by the British Civil Service in 2013 to guide its staff on how to write online policy documents. Such documents are designed to deal with the day-to-day implementation of the policy decisions of government. The 2013 Style Guide caught the attention of the media because of a section on "plain English", which contained words to avoid, or "banned words" as they became to be referred to in the press. One of the points that Philip's analysis illustrates is how meanings of words depend on both their phraseological and genre context. At the same time, Philip's chapter represents a specific, highly specialised type of discourse, which can also be conceived as a *metadiscourse* of the Civil Service discourse.

An important aspect of the discourse that corpus linguistics can help to investigate is that meaning is not only a monolingual matter. Meanings can be compared across languages and translated from one language into another. Teubert stresses: "Meaning is the core issue of translation. A translator produces a paraphrase of a text in another language" (Teubert 2002: 190–191). In this volume, Aijmer presents a classic corpus-based contrastive study of two cognate words – English *must* and Swedish *måste* – to show how they differ both formally and functionally. Using a bidirectional English-Swedish parallel corpus of fiction, she illustrates how translation equivalence can be used as a criterion for sameness of

meaning. The contribution by Granger and our own chapter also deal with meanings across languages.

Methods

The way in which we approach methods in this book is from an analytical point of view. We have deliberately not included chapters that focus on the introduction of new software, corpus formats or interfaces. This is not to say that such topics are not important in corpus linguistics, far from it. But our aim is to concentrate on the concepts that contribute to our corpus linguistic understanding of the discourse. Patterns that shape the meanings of words and that are a reflection of repeated usages play an important role for quantitative methods to capture discourse phenomena. Such corpus approaches emphasise the "incremental effect of discourse" (Baker 2006: 13): repeatedly occurring patterns of language use create discursive representations of people, events, intuitions, etc. As corpus methods make it possible to trace such patterns, corpus studies are increasingly used to study the representation of social issues, global events, groups in society, or what might sometimes be called 'cultural key words'. For instance, Stubbs (1996: 176ff.) discusses examples from the lexical field of WORK, Jaworska and Krishnamurthy (2012) investigate notions of *feminism*, Teubert and Čermáková (2004) look at *globalisation*, and Alexander (2002) and Mahlberg (2007) study *sustainable development*. To identify meanings in different discourses, it is helpful to begin with frequency comparisons. A commonly used method is the keywords procedure. Grundmann and Scott (2012), for instance, draw on key comparisons to study climate change discourse by comparing news coverage across four different countries. In the present volume, Partington and Zuccato, Baker as well as Granger also include key comparisons among the methods they employ in their studies.

Both Stubbs and Krishnamurthy highlight how the focus on the computer-supported aspects of corpus linguistics provides a far too narrow picture. Their discussion of the use of concordances stresses that the display technology is mainly a way to visualise data in support of an analysis. Sinclair's (2003) introduction to concordance analysis also shows this point clearly. Importantly, the title of his book *Reading concordances* reflects the qualitative input required by the analyst. The keywords procedure might appear to be a more objective method than a concordance analysis because it involves statistical tests to identify keywords – compared to a concordance analysis where the researcher identifies patterns. However, the critically reflective approach that Baker takes to this common corpus linguistic method demonstrates the limits of objectivity. He emphasises that simply generating a keyword list does not constitute an analysis. The researcher will

have to decide on which keywords to focus for a more detailed analysis. As Baker is interested in the extent to which a keyword analysis reduces bias – or cannot avoid a certain amount of it – he conducts a reflexive analysis of six of his own keyword studies.

Connections

The papers in this volume show that the concepts of 'meaning' and 'methods' are not only fundamental to corpus linguistics, they also allow us to see connections between different fields and ideas. For corpus linguistic approaches that deal with more than one language connections to contrastive linguistics or translation studies are almost unavoidable, as Aijmer shows in her chapter when she draws on the notion of 'translation equivalence'. Granger explicitly makes connections between fields by proposing a methodological framework that she terms 'Contrastive Translation Analysis'. This framework aims to shed fresh light on the notion of the 'third code' or translated language. Granger combines not only corpus linguistics and translation studies but also draws on learner corpus research to identify distinctive features of translated language.

In his chapter, Stubbs pointedly reminds us that "the reason why many precursors of important ideas sank without trace is that they were not connected to other ideas". Barlow's chapter is a good example to illustrate how corpus linguistic questions can make connections across linguistic disciplines that do not easily seem to go together. The strengths of corpus linguistics are in its focus on the identification of tendencies and patterns of mainstream language use. The more repetitions we find of patterns and meanings the clearer the picture becomes. Barlow's concern with individual variation is thus a rather innovative topic for corpus linguists, and he tackles it by beginning to make links to long-standing concerns in linguistics. Barlow is interested in individual variation as a baseline for assessing corpus results based on aggregate data. To gain a corpus linguistic understanding of the connections between the individual and the group he draws on research from socio- and cognitive linguistics and touches on recurrent ideas and dilemmas within linguistics.

Teubert often argues that corpus linguistics and cognitive linguistics are rather separate or even incompatible strands. For Teubert language is "what takes place between people, not inside them" (Teubert 2010: 114). As a consequence, corpus data is all the evidence that is needed to describe meanings. With the distinction that Barlow proposes between a comprehension and a production grammar, this approach to corpus evidence is called into question. With our own chapter on literary translation we also raise related questions. Using the examples of Alice

and the Queen from *Alice's Adventures in Wonderland*, we argue that the notion of mind-modelling (Stockwell 2009) can account for process-, product- and function-oriented aspects of literary translation. Thus we combine corpus linguistic evidence of meanings with a cognitive approach that emphasises individual meanings that are created in the mind of the reader – or the translator as a reader. In the study of literary texts, such individual meanings are also referred to as readings. The fact that Sinclair (2003) similarly talks about "reading" concordances is a worthwhile connection to make in this regard and goes some way to explaining why the analysis of concordances is not a fully automated process and different "readers" come up with different meanings derived from concordances. These differences might be less striking than the differences between readings of a literary texts. Based on the cumulative evidence a concordance provides we would expect rather converging readings. However, the concept of the semantic prosody is an example that illustrates how individual variation can affect the reading of concordances. The semantic prosody is the most fuzzy category of the lexical item (Sinclair 2004). Unlike collocations, which are based on the verbatim repetition of words in a concordance, semantic prosodies characterise attitudinal or evaluative meanings and are manifested in a variety of forms. Hence, their identification is more open to differing readings.

This book is a contribution to the discourse of corpus linguistics, where researchers propose new approaches and reflect on the development of corpus linguistics, both in terms of its history and in relation to other disciplines. The samples from the corpus linguistics discourse are, however, selective. Our aim has not been to capture *the* discourse. What the book highlights is how we, as corpus linguists, negotiate the meanings of the central concepts and frameworks we draw on. The book is a tribute to Wolfgang Teubert and his contribution to our field and the study of discourse. It shows some of the traces he and many others have left in the discourse of corpus linguistics.

References

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

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. & Ellece, S. 2011. *Key Terms in Discourse Analysis*. London: Bloomsbury.

Brown, G. & Yule, G. 1983. *Discourse Analysis*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511805226

Grundmann, R. & Scott, M. 2012. Disputed climate science in the media: Do countries matter? *Public Understanding of Science* 23 (2): 220–235. https://doi.org/org/10.1177/0963662512467732

Jaworska, S. & Krishnamurthy, R. 2012. On the F word: A corpus-based analysis of the media representation of feminism in British and German press discourse, 1990–2009. *Discourse & Society* 23(4): 401–431. https://doi.org/10.1177/0957926512441113

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

Partington, A. 2010. Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora* 5(2): 83–108. https://doi.org/10.3366/cor.2010.0101

Sampson, G. 2013. The empirical trend: Ten years on. *International Journal of Corpus Linguistics* 18(2): 281–289. https://doi.org/10.1075/ijcl.18.2.05sam

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

Sinclair, J. M. 2003. *Reading Concordances*. London: Pearson/Longman.

Sinclair, J. M. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stockwell, P. 2009. *Texture: A Cognitive Aesthetics of Reading*. Edinburgh: Edinburgh University Press.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Tannen, D., Hamilton, H. E. & Schiffrin, D. (eds). 2015. *The Handbook of Discourse Analysis*, 2nd ed. Oxford: Wiley Blackwell.

Teubert, W. 2001. A province of a federal superstate, ruled by an unelected bureaucracy. Keywords of the Euro-sceptic discourse in Britain. In *Attitudes Towards Europe*, C. Good, A. Musolff, P. Points & R. Wittlinger (eds), 45–88. Abingdon: Ashgate.

Teubert, W. 2002. The role of parallel corpora in translation and multilingual lexicography. In *Lexis in Contrast: Corpus-based Approaches* [Studies in Corpus Linguistics 7], B. Altenberg & S. Granger (eds), 189–214.Amsterdam: John Benjamins. https://doi.org/10.1075/scl.7.14teu

Teubert, W. 2004. Language and corpus linguistics. In *Lexicology and Corpus Linguistics. An Introduction*, M. A. K. Halliday, W. Teubert, C. Yallop & A. Čermáková, 73–112. London: Continuum.

Teubert, W. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13.

Teubert, W. 2010. *Meaning, Discourse and Society*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511770852

Teubert, W. & Čermáková, A. 2004. Directions in corpus linguistics. In *Lexicology and Corpus Linguistics. An Introduction*, M. A. K. Halliday, W. Teubert, C. Yallop & A. Čermáková, 113–165. London: Continuum.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Viana, V., Zyngier, S. & Barnbrook, G. 2011. *Perspectives on Corpus Linguistics* [Studies in Corpus Linguistics 48]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.48

Worlock Pope, C. (ed.). 2010. The Bootcamp Discourse and Beyond. Special issue of *International Journal of Corpus Linguistics* 15(3).

The (very) long history of corpora, concordances, collocations and all that

Michael Stubbs
University of Trier

In the development of academic disciplines, important ideas are often proposed, forgotten, and then rediscovered much later, when they are connected to other ideas in a way which reveals their significance. I give examples of ideas which are often thought of as quite modern, although they have a very long history:
– using corpora in constructing dictionaries and language teaching materials
– using concordances as data for textual exegesis and information retrieval
– using collocations as evidence of word meaning.
In all three cases the theoretical significance of the ideas became clear only after improved techniques of visualisation allowed patterns to be seen in complex non-numerical data.

1. Overview

The many accounts of the history of corpus study are all partial in different ways, and my account here is also a mere sketch of the origins of some major ideas. My main focus is on the history of attempts to visualise patterns in texts and how these attempts led, over a long period of time, to important linguistic concepts.[1]

First, important ideas are often proposed, forgotten about, and then rediscovered, independently, sometimes hundreds of years later. In the natural sciences, so it is said, "science destroys its past" (Kuhn 1969). But a survey of computing in the humanities makes the point that:

> It is good every so often to remind ourselves that we are part of a line of historical development. It is particularly appropriate to do this in times which are described as new, ground-breaking, unique … [W]e may […] labour under the misconception that what we are doing is so radically different from before. (Fraser 1996)

1. For some (non-British?) readers I should perhaps point out that the title of my article alludes to two books: Sinclair (1991) and Sellar and Yeatman (1931).

https://doi.org/10.1075/scl.87.02stu

Second, the reason why many precursors of important ideas sank without trace is that they were not connected to other ideas, and so their full implications could not be seen. This process of connection sometimes took hundreds of years. Third, corpus linguistics, in its modern sense, could obviously develop only with computational help. However, it is not technology alone which leads to theory, but the techniques of visualisation which the technology makes possible, and the importance of visualisation was recognised long before computers were available. Fourth, practice came before theory. Real world problems required practical applications which, in turn, led to theoretical developments (often helped by visualisation techniques). Many important concepts originated in textual exegesis, language teaching, lexicography, and information processing.

2. Previous work

We have no comprehensive history of corpus study and its central ideas, including concordance and collocation, and no comprehensive bibliography. Much of the chronology is, however, well documented. Francis (1992) and Meyer (2008) provide good coverage of pre-electronic corpora. Kennedy (1998: 13–87) covers both pre-electronic corpora and concordances, and the design and mark-up of digital corpora. Renouf (2007) reviews digital corpora from small beginnings in the 1960s, via multi-million-word corpora in the 1990s, to the use of the world-wide-web as corpus in the late 1990s. Fraser (1996) and Hockey (2004) are invaluable sources on the chronology, post-1940, of scholars, publications, events, technological developments, and the essential institutional framework. They document the first journals, conferences and university centres (in the 1960s), increasingly user-friendly software (from the 1960s onwards), and increasingly coordinated work (e.g. see Sperberg-McQueen and Burnard (1990) on the *Text Encoding Initiative*). Their factual accounts provide the essential external history of events, with brief details of applications (e.g. in biblical exegesis and authorship attribution), but contain only limited evaluation of the ideas.

Accounts by scholars who were themselves involved in the early days of computer-assisted work can be particularly fascinating, but they must be read as autobiographical interpretations by major participants. Leech (1991) and Francis (1992) provide accounts from the early 1990s, and Busa (2004), Sinclair (2004), Svartvik (2007), Johansson (2008) and Leech (2013) provide later retrospectives. In an ironic account of the reception of his own work – implicit in his highly evaluative vocabulary – Sinclair (1991: 1) identifies a succinct historical pattern:

> Thirty years ago [in the 1960s] when this research started it was considered impossible to process texts of several million words in length. Twenty years ago [in the 1970s] it was considered marginally possible but lunatic. Ten years ago [in the 1980s] it was considered quite possible but still lunatic. Today [in the 1990s] it is very popular.

Teubert (2004: 107–112), in a "brief history" of corpus linguistics, emphasises a further pattern within this pattern. Much empirical descriptive work from the 1970s to the 1990s emphasised lexis and grammar, whereas "the quest for meaning all but disappeared from the agenda". Like everyone else, Teubert has his own agenda, which is clear from his citation of British work on collocation by Palmer and Hornby in the 1930s, by Firth in the 1950s, and by Sinclair in the 1960s.

From her French perspective, Léon (2005) questions the selectivity of an orthodox "common history" which scholars have retrospectively built, and argues that it leaves serious gaps and makes exaggerated claims about "a new philosophical approach" to language study. At a factual level, she questions the frequent claim that the *Brown corpus* (Francis & Kučera 1964) had no precursor, pointing out that the digitisation of the *Trésor de la langue française* had begun before the *Brown corpus* was planned. (She herself fails to record that Busa's digital corpus was begun in 1947, see below.) At a theoretical level, she questions several frequent claims: that there was a simple rationalist-empiricist opposition between Chomskyans and corpus linguists, that Chomsky's work interrupted corpus-based work in the 1960s and 1970s, and that corpus linguistics can claim to be an autonomous area.

In an influential article on the historiography of linguistics, Hüllen (1996: § 79) emphasises that there is no objective memory and no objective reconstruction. My own reconstruction here will emphasise the Firthian insistence on empirical methods of semantic analysis.

3. Concordancing content

Early language study had largely religious aims:

> The first philologists and the first linguists were always and everywhere priests. History knows no nation whose sacred writings or oral tradition were not to some degree in a language foreign and incomprehensible to the profane. To decipher the mystery of sacred words was the task meant to be carried out by the priest-philologists. (Voloshinov 1929[1973]: 74)

Around 1700 years ago, one such priest-philologist, Eusebius of Caesarea, proposed a method of systematically comparing related texts. He is probably

ultimately responsible for the term 'concordance', though what he produced was not a concordance in exactly the modern sense.

Eusebius was a Roman scholar who became Bishop of Caesarea in Palestine around 314, and wrote widely on biblical criticism and church history. He deserves a place in a history of text analysis because he developed a system which makes it easier to compare the New Testament accounts by the four Evangelists. In a letter to a Christian called Carpianus, Eusebius (c. 320?) explains that he had improved a system devised around 220 by Ammonius of Alexandria in his *Harmony of the Gospels* (which is now lost). Eusebius' aim was to show the *concordantia*, the agreement or harmony, between the four canonical Gospels.

He divided the four texts into numbered sections, so as to provide a system of cross-references which allowed a reader to find and compare parallel passages. He then designed an index in the form of ten Tables which show these textual parallels: that is, sections where each Gospel had the same or different content from the other three. In parallel columns are the numbers of the sections common to all four gospels, or three, or two, or those unique to each evangelist.[2] This cross-reference system is known as the 'Canon Tables', and it provides a simple but precise representation of intertextual relations, which shows, as we say, "at a glance" where the Gospels agree or differ. In some of the most famous representations (such as the *Book of Kells*, c. 800 CE; Meehan 1994), the visualisation is certainly beautiful, but the decoration is so elaborate that it makes the information almost unreadable. Many other more useable examples can be seen in the catalogue of illuminated manuscripts in the British Library (see Figure 1).[3]

An obvious difference between Eusebius' Tables and a concordance in the usual modern sense is that the Tables are based on a content analysis, and are therefore independent of the language (Latin, English or whatever) into which the texts are translated. In addition, a tabulation of this kind is, of course, only applicable to different versions of a text which all have very similar content.

The Canon Tables were used up to and throughout the Middle Ages (Oliver 1959). There was then a gap of several hundred years before related ideas were developed, and there is no clear causal connection between Eusebius' work and later concordances. However, ideas, once expressed (in writing), remain in the intellectual world and can be taken up at any future time. In modern terms, one could say that Eusebius recognised that texts can be turned into data in a way which allows patterns to be visualised. He transformed texts into numerical grids in order to facilitate systematic comparison, which is an essential prerequisite of any textual criticism.

---

2. Eusebius does not explain why two comparisons are missing: Mark/Luke/John and Mark/John.

3. At <http://www.bl.uk/catalogues/illuminatedmanuscripts/searchSimple.asp> search for "canon tables" (17 September 2018).

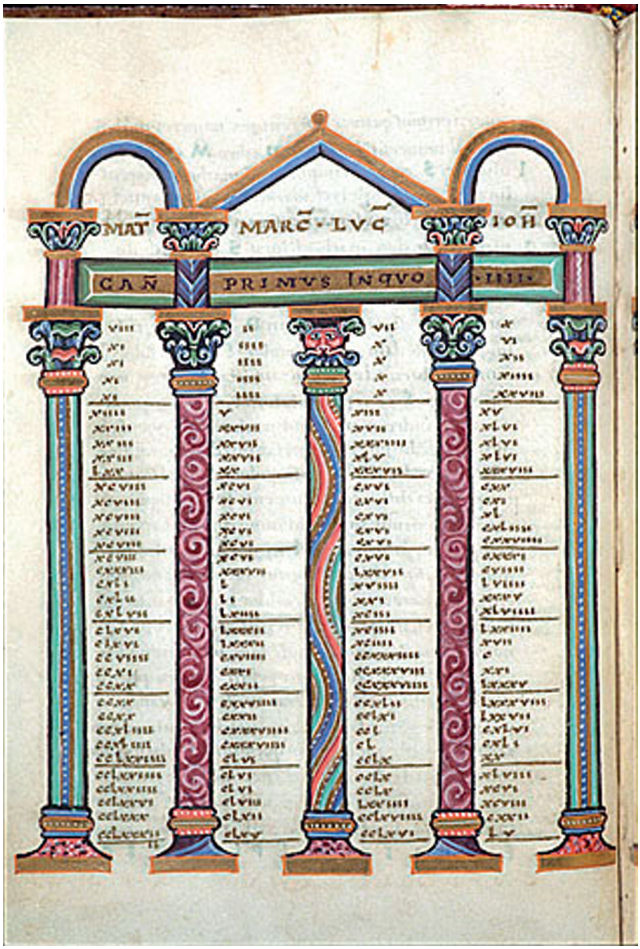

Figure 1. Canon Tables, Mid-1200s, Echternach, Luxembourg
This image identified by The British Library is free of known copyright restrictions.

4. Index, verbal concordance and "real" concordance

Since a concordance, in the usual modern sense, tracks lexis across texts, translations of a text in different languages or different translations in the same language result in different concordances (unlike Eusebius' Tables). However, terms such as 'index' and 'concordance' are often not clearly distinguished. Even the 1933 edition of the *Oxford English Dictionary* (OED) is not very clear. It defines 'concordance' as:

> An alphabetical arrangement of the principal words contained in a book, with citations of the passages in which they occur. These were first made for the Bible; hence Johnson's explanation *A book which shows in how many texts of scripture any word occurs*. (p. 773)

This does not clearly distinguish between a list of "principal" (= content/lexical?) words with their places of occurrence (an index?) and a list of words together with their surrounding short co-text (a concordance?). It then adds a distinction between a "verbal concordance" as distinguished from "a real concordance which is an index of subjects or topics." This definition would make Eusebius' Tables a "real concordance".

Early "verbal concordances" of the Bible were produced from the Middle Ages onwards. Liberman (2004) provides some details of the first recorded usages of the word in the OED (1387, "a greet concordaunce uppon þe bible", and later citations up to the 1800s), and also cites information from the *Catholic Encyclopedia* and the *Jewish Encyclopedia*. In 1230, Hugo de Sancto Caro (Hugh of St Cher), working in Paris with 500 Dominican monks, produced an index of words in the Latin Vulgate translation. Initially, this listed only where a word occurs. Between 1250 and 1252, three English Dominicans added the complete quotations around the occurrences listed, and therefore turned it into a concordance. This, in turn, contained only the content words, and around 1435 another Dominican, John of Ragusa, added the grammatical words (e.g. *nisi, ex, per*). Liberman (2004) comments that Hugo must have been working with a very inefficient algorithm if he required 500 monks. After all, in the mid-1700s, Alexander Cruden produced a concordance of the Bible single-handedly in only twelve years (while employed to do other work).

5. Concordancing form

In the Preface to his own *Complete Concordance to the Holy Scriptures*, Cruden (1737) lists the index by Hugo and concordances in several other languages, including Latin, Hebrew, Greek, English, French and Dutch.[4] Cruden was born into a Scottish Calvinist family who held the Presbyterian belief that people should study the Bible in their own native language. His own severe education required him to learn by heart many passages of the Bible, and therefore to pay close attention to the text. He appears also to have held the fundamentalist belief that the Bible expressed the literal truth in God's own words (Keay 2004: 168). These two beliefs seem logically contradictory: at the very least, they depend on an optimistic faith in the possibility and accuracy of word-by-word translation. His intention to train for the ministry was made impossible, for reasons explained by Keay (2004), and as a result he saw his life's work in constructing the *Concordance*, as an aid to Bible study. He concordanced the 1611 "Authorised Version" of the English Bible

4. The *Concordance* is available from <http://books.google.de>

commissioned by King James I. For each content word, his *Concordance* gives its position in the whole text, and six or seven words of co-text to left and/or right.

The Preface states explicitly his view on the relation between meaning and use:

> A CONCORDANCE is a Dictionary, or an Index to the Bible, wherein all the words used […] are ranged alphabetically, and the various places where they occur are referred to, to assist us in […] comparing the several significations of the same word.

Cruden's terminology is rather muddled, though he clearly understood that his concordance lines provided systematic and observable evidence for the textual meanings of words (which might be different from their usual dictionary meanings). Indeed, he is confident that his *Concordance* is an improvement over earlier work. It allows one passage to be compared with another, which is "the best rule of interpreting Scripture", and thereby shows "the various significations" of words, since it allows many important things to be "observed at one view". He does however recognise the limitations of his less than ideal layout, and the problems imposed by the huge size of the book, the expense of paper, the unreadability of small fonts, and so on (see Figure 2).

Cruden has a rough concept of phraseology and collocation, and sometimes lists phrases and collocations separately from individual words: for example, not only *dark* and *darkness*, but also *land of darkness*, *in darkness*, etc., and *darkness* with *day* and *darkness* with *light*. And he lists separate meanings of *darkness*, such as "want of natural light", "hell", "ignorance", etc.

His theoretical comments are somewhat random. For example, the entry for DRY observes that "by the words annexed to DRY the meaning is obvious", but of course the difficulty comes in making explicit what seems "obvious". However, he was not centrally concerned with analysing or interpreting the data which could be observed with a concordance, but with the essentially practical aim of designing an information retrieval system for preachers preparing their sermons, and in this he was very successful. He worked extremely accurately: his results have been checked against computer-assisted searches by Barnbrook, Mason and Krishnamurthy (2013: 213). On the criterion of precision, his work rates as excellent: as one would expect with a manual search, everything he finds is relevant. On the criterion of recall, it rates nearly as well: he finds almost everything which is relevant. However, he does for example miss a relevant example of *light* collocating with *darkness*. Compare Figure 2 with this example:

> *Genesis* 1:5. And God called the *light* Day, and the *darkness* he called Night.

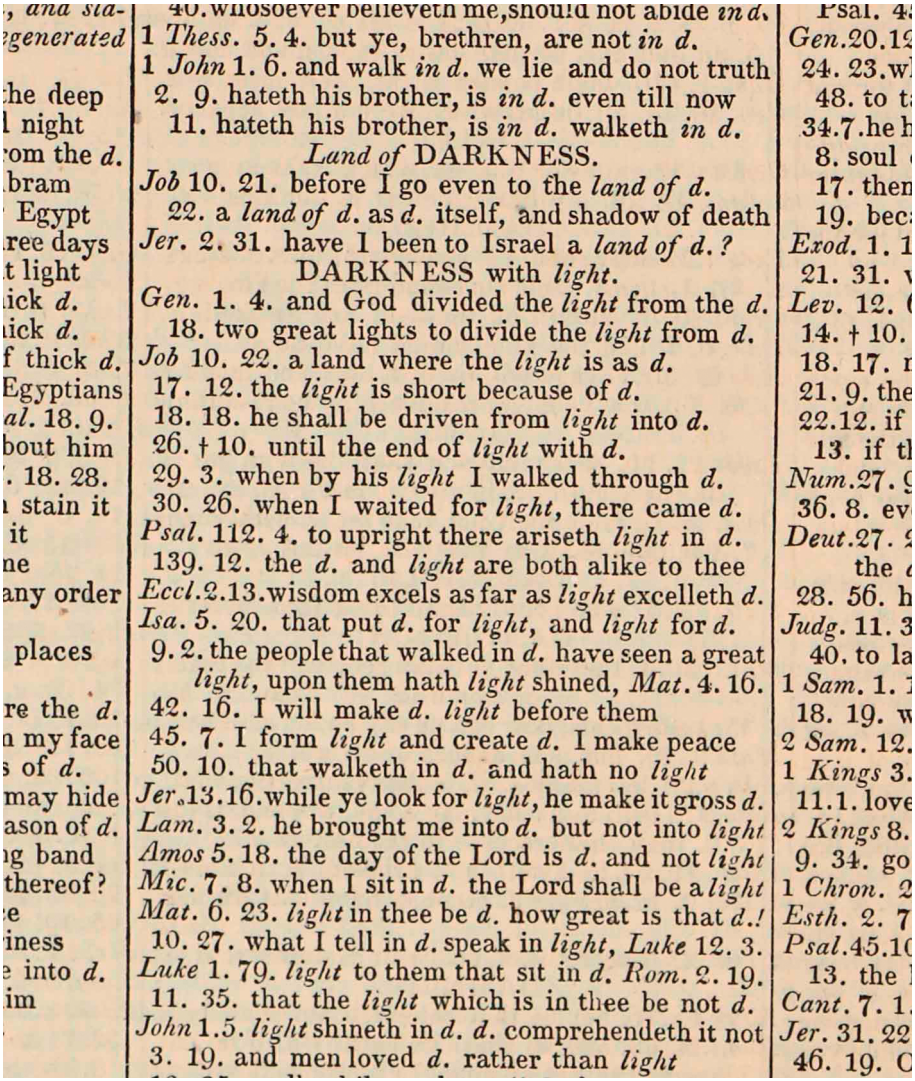

Figure 2. A page from Cruden's *Concordance* of 1737
This photograph is of the tenth edition of 1833.
Photograph M. Stubbs

Other examples from his work and its interest for corpus linguistics are discussed by Kennedy (1998: 13–14) and Barnbrook, Mason and Krishnamurthy (2013: 96–99, 207–13).

Cruden produced two further editions of his *Concordance* with improved layouts. Over sixty editions (complete, abridged, etc.) have appeared since his death, and it has never been out of print. It did not take long before concordances

were also produced of major literary works. Cruden (1741) himself produced a *Verbal Index to Milton's Paradise Lost*. And Samuel Ayscough published *An Index to the Remarkable Passages and Words made use of by Shakespeare; Calculated to Point out the Different Meanings to which the Words are Applied*. This is in fact a concordance, although only of selected node words, and often in only three or four words of co-text (Ayscough 1790).

Both indexes and concordances are estrangement devices designed to explore texts and to alter our view of language. Indexes locate passages and help to identify parallels. Concordances go much further: they rip texts apart and create new textual objects from fragments of the original, in which "many important things may be observed at one view", as Cruden puts it. This point is made implicitly in the sub-title of Cruden's biography, which talks of the genius who "unwrote the Bible" (Keay 2004). Both Cruden (in his Preface) and Ayscough (in his title) state a theory which is often attributed well over a hundred years later to Wittgenstein (1953) and Austin (1962), that "meaning is use": "The meaning of a word is its use in the language" (Wittgenstein 1953: § 43).

6. Meaning and use

Samuel Johnson was working at around the same time as Cruden, and expressed similar ideas about the relation between meaning and use in his *Dictionary of the English Language* (1755). Its title page advertises a book "in which the words are deduced from their originals, and illustrated in their different significations by examples from the best writers". Whatever we think of his restricted corpus of texts by "writers before the Restoration" (i.e. 1660, almost 100 years before the *Dictionary*), Johnson's was the first English dictionary to use empirical textual data as evidence of word meaning.

It is probably impossible to fully appreciate how primary sources from over 200 years ago would have been read at the time. To understand them as interventions in cultural debates would require detailed knowledge of other contemporary texts and of social and moral assumptions which were unquestioned in his time. Johnson is nowadays usually quoted – and made fun of – for his conservative views of both his corpus and his aims. In the Plan for his *Dictionary*, Johnson (1747) asserts that "all change is of itself an evil", that he will aim to "fix the English language", and that in choosing his data he will prefer "writers of the first reputation".

However, he admits a serious problem in selecting authoritative data ("who shall judge the judges?"), and he also has precise non-prescriptive ideas about language varieties and lexical organisation. One difficulty is defining core vocabulary: for everyday words (e.g. *horse, dog, cat*) "it will be hard to give an explanation,

not more obscure than the word itself". Another difficulty is how much specialist vocabulary to include, for example, the "terms of science" and "the peculiar words of every profession" such as "law, merchandise, and mechanical trades". He knows that collocations are sometimes arbitrary: *he died of his wounds*, but *he perished with hunger*. He knows that "phraseology" is important: he gives many examples, such as *to make way, to make a bed, to make merry*. And he appreciates the effect of collocation on meaning, in examples such as:

1. *The bird fell to the ground.*
2. *The silk had blue flowers on a red ground.*
3. *The ground of his work was his father's manuscript.*

In the Preface to the *Dictionary*, he rails prescriptively against "spots of barbarity impressed so deep […] that criticism can never wash them away", and against working class language: the "fugitive cant" of "the laborious and mercantile part of the people" which is "unworthy of preservation". However, he knows that language does constantly change and that there are many different text-types, even if he "could not visit caverns to learn the miner's language". And he expresses, if not very explicitly, principles which are still central to modern lexicography. He understands how the meaning of a word can be determined by co-occurring words, and he states a principle still followed by corpus lexicographers in choosing examples to illustrate word meaning:

> It is not sufficient that a word is found, unless it be so combined as that its meaning is apparently determined by the tract and tenour of the sentence.
>
> (Johnson 1755)

The relations between meaning and use became clearer in the late 1800s in the most famous English dictionary of all, the *Oxford English Dictionary* (OED). Its production required the lexicographers first to construct a diachronic corpus, so that they could then give an account of the history of the language by listing chronologically examples of the language itself. A full account of the textual materials, the design of the data-base, and the role of its first editor, James Murray, are discussed comprehensively elsewhere (Murray 1977, Winchester 1999, 2003; Mugglestone 2005, Brewer 2011, http://www.oed.com/).

It is worth noting that it was another priest-philologist, the Anglican archbishop Trench, who first proposed the OED, in a paper to the *Philological Society in London* (Trench 1857). Trench is also discussed in detail elsewhere (Crowley 1989: 51–90, Rastall 2001).

7. Practice and theory

Scholars were initially interested in texts which had unique religious, philosophical and/or literary value, and only much later in diverse texts from everyday life. Samuel Johnson (1755) "studiously endeavoured to collect examples and authorities from the writers before the Restoration", but he also recognised that "words must be sought where they are used".

As noted above, pre-electronic corpora are covered well by Francis (1992) and Meyer (2008). However, a brief list emphasises that ideas do not develop in a social vacuum, and that most of the corpora which were set up from the late 1800s onwards had very practical aims (see also Howatt 2004: 264–352).

- Kaeding (1897) used a corpus of over 11 million running words of written data to produce the first detailed German frequency count (of letters, syllables and words) and thereby improve shorthand techniques.
- Thorndike (e.g. Thorndike & Lorge 1944) made word-frequency lists from corpora of up to 18 million running words (still with a strong literary bias) in order to improve literacy materials for children and adults.
- West (1953) used a corpus of 5-million words to produce a "general service" list of around 2,000 words of maximum utility to learners of English. It recorded both word frequencies and also the relative frequencies of different word-senses.
- Gougenheim et al. (1956) used a corpus of 312,000 running words from 163 audio-recorded conversations to construct materials for teaching French as a foreign language: *Le français fondamental*.

Jespersen and Fries are mainly known for their descriptive/theoretical grammars (Jespersen 1909 [1949]; Fries, C. C. 1952), but both had lifelong interests in language teaching. Jespersen collected and stored thousands of quotations in shoeboxes and in pigeon holes in his desk. Fries recorded 250,000 words of telephone conversations from 300 speakers (for detailed discussion see Fries, P. H. 2012).

The *Survey of English Usage*, founded in 1959 by Quirk, was at the turning point between manual and computer-assisted analysis. Samples of naturally-occurring written and spoken British English were initially stored and annotated on paper cards, and were then, from the late 1970s onwards, digitised and analysed with software tools (see http://www.ucl.ac.uk/english-usage/ (17 September 2018)).

8. Digital corpora

The first computer-assisted corpus analysis had started earlier, but initially remained unknown to many linguists. This was work by Roberto Busa (1913 –2011), an Italian Jesuit priest.[5] Busa (2004) recounts how, in the late 1940s, he persuaded IBM to start entering the complete works of Thomas Aquinas into a computer, and then (with some self-irony) how the project developed. The initial version of the *Index Thomisticus*, would have required 13 million punch cards, one for each word, with a context of 12 lines stamped on the back. However, "in His mercy, around 1955, God led men to invent magnetic tapes", which simplified things, though "their combined length was 1,500 km, the distance from Paris to Lisbon". The printed version of the *Index* comprises 20 million lines on 65,000 pages of 56 volumes. A CD-ROM ("cum hypertextibus") appeared in 1992 (Busa 1974 [1992]). An on-line version with a powerful search-engine became available in 2005. Work on the data-base is still (2014) continuing with the aim of syntactically annotating the entire corpus to form a dependency-based treebank.[6]

Busa (1976, 1992) demands more of computers than their mere ability to carry out routine secretarial tasks quickly and efficiently. Too often, he says (Busa 1992: 128), the computer "is used to reach the same targets as before, using the same methods as before. Philologists must create new strategies for new goals, when using computers". He goes on to list many such tasks for a "new philology". Ramsay (2008) notes the radical hermeneutic implications of Busa's work for a new way of visualising textual patterns:

> The founding moment [of the digital humanities] was [Busa's] creation of a radically transformed, reordered, disassembled and reassembled version of one of the world's most influential philosophies. (Ramsay 2008)

5. For example, Busa's work is not mentioned by Kennedy (1998) or Barnbrook, Mason and Krishnamurthy (2013). I am not sure what to make of this. Was it not known? Not regarded as relevant? Because it was done outside the English-speaking world and on Latin to boot? Because it was not reported in the right conferences and journals? If it was not known then, logically, it could not have influenced later work.

6. The main search page at <http://www.corpusthomisticum.org/it/index.age> has a detailed description in English. The completed part of the *Index Thomisticus Treebank* is available at <http://itreebank.marginalia.it/> (17 September 2018). I am grateful to Marco Passarotti of the Università Cattolica del Sacro Cuore, Milan, for this update.

9. Collocation and phraseology

Arguably the most far-reaching discovery of computer-assisted corpus analysis is that it is not individual words, but recurrent collocations, which are the basic units of meaning. This idea, like many others, had its origins much earlier and developed only slowly.

Cruden and Johnson had elementary ideas about phraseology and collocation. In work on French, Bally (1909) pointed out that conceptual units are often idiomatic multi-word phraseological units. He talks of "groupements usuels" and "locutions phraséologiques". In work on German, Porzig (1934) discusses cases where one word implies another. Examples with verbs include: *lick* implies *tongue*, *kiss* implies *lips*, *bite* implies *teeth*, etc. This would predict that, if such a verb occurs, the corresponding noun will usually not co-occur (?*she kissed him with her lips*). But such a hypothesis can be investigated only with large corpora of naturally occurring data, to which Porzig had no access.

A very substantial report on collocation was prepared by Palmer (1933) following his language teaching experience in Japan. A 20-page introduction makes explicit many points which are still central to semantic theory. First, a collocation is "a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts". Second, word frequency lists are misleading, because individual words occur with different meanings in different phraseology (compare *fork* and *fork in the road*, Palmer 1933: 12). Third, collocations provide different problems for comprehension and production. This principle underlies modern corpus-based dictionaries, which aim to help learners both to understand and use phraseology idiomatically. Palmer then lists thousands of collocations: his classification is not intended for learners of English, but "composed by technicians for technicians" for those preparing dictionaries or teaching materials.

The idea that the meaning of a word depends on its typical collocates was proposed more or less clearly by Cruden, Johnson, Palmer and others. Although attempts at machine translation were widely criticised and rejected by linguists, in a paper on the topic, Weaver (1955) makes the idea even more explicit and explains very simply what he means by "statistical semantics":

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. *Fast* may mean 'rapid'; or it may mean 'motionless'; and there is no way of telling which. But if one lengthens the slit in the opaque mask, until one can see not only the central word in question, but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word. […] The practical question is, what minimum

> value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word? [Cited from a 1949 version]

At around the same time, in work which is usually taken as the starting point for statistical work on collocation, Firth said even more simply: "you shall know a word by the company it keeps" (Firth 1957: 11). This is the maxim known to lawyers as "noscitur a sociis" (it is known by its associates). That is, the meaning of an ambiguous word is to be determined by the words surrounding it.[7]

10. Meaningful quantification

Firth's formulation assumes, however, that it is individual words which are the basic units of meaning. Still several years before corpus resources became generally available, McIntosh (1961) made the next move in the argument, and set out major principles which have been thoroughly documented by corpus study. Not only do the collocates of a word go "a long way towards constituting [its] meaning", but also "the lexical item and the word are not co-extensive".[8]

The *OSTI Report* (Sinclair, Jones & Daley 1970[2004]; UK Government Office for Scientific and Technical Information) describes research carried out between 1963 and 1969 on 135,000 words of computer-readable spoken text. It is seminal work, but was hardly available for over thirty years until it was formally published in 2004. The main aim was to develop a quantitative theory of collocation by discovering the relation "between statistically defined units of lexis and postulated units of meaning" (Sinclair 2004: 6). The research made substantial progress with a question which has still not had a satisfactory answer: How can the units of meaning of a language be objectively identified? It is important to emphasise that this tradition of text-based work, from Firth to Sinclair, was concerned, from the beginning, with a theory of meaning.

The way in which semantic theory then rapidly developed out of practical lexicography, when work on the COBUILD series of dictionaries and grammars began in the 1980s, is thoroughly documented elsewhere (e.g. Sinclair 1987, Moon

7. Dorothea Halbe has pointed out to me that Firth's phrasing is similar to that in Henry Fielding's comic novel *Tom Jones* (1749): "… that short Latin proverb, 'Noscitur a socio' which, I think, is thus expressed in English, 'You may know him by the company he keeps.'" I had always assumed that the phrase originated with Firth, and am grateful for her correction.

8. McIntosh's work in military intelligence had influenced his thinking about quantitative empirical research on language. At the University of Edinburgh, where he worked with Halliday and Sinclair, he witnessed the beginnings of modern computing, and was a founding member of the School of Epistemics (Giegerich 2005).

2007). Hanks (2013) discusses in detail the kind of empirical corpus-based lexicography which developed in the wake of the COBUILD project.

A second early computer-assisted project shows how easy it is to underestimate the long history of important ideas. In 1965, Sture Allén set up the research group *Språkdata* at the University of Göteborg in Sweden, and he and his colleagues published a Swedish frequency dictionary based on a one-million-word newspaper corpus (Allén et al. 1975). The introduction notes the intellectual climate in the 1970s, when "linguists tended to lay too much emphasis on introspection" (Allén et al. 1975: xxxi), and emphasises the "meaningful quantification" of authentic material. It also discusses several methodological, conceptual and cognitive principles: recurrence as "the methodological foundation of the investigation" (p. xxxiii), phraseology as "the area of intersection between grammar and lexicon" (p. xxxii), and the pervasive nature of "recurrent blocks" as a central part of "a realistic, psycholinguistically plausible model" of language use (p. xlii). In addition, the authors propose a collocational measure (p. xlv): the "constructional tendency" of a word is the ratio of its overall frequency to its frequency in a recurrent string. To take an English example, the lemma QUESTION occurs over 42,000 times in the *British National Corpus*. The word *vexed* occurs only 165 times, but when it does occur, it has nearly a one-in-three chance of occurring in the 2-gram *vexed question(s)*. Many current measures of collocational attraction still do not deal with such radical asymmetry.[9]

Other early work, for example papers presented at a conference in 1970 (Wisbey 1971), shows the high conceptual standard which was possible with software which from today's perspective seems very cumbersome.

11. KWIC (Key word in context) concordances

A technique which could make recurrent collocations visible "at a glance" had itself been developed for quite different purposes over a long period of time. Before phraseology could be systematically studied in large data collections, a practical invention had to be combined with a theoretical idea.

The practical invention was punch cards, which represent information digitally by the presence or absence of holes in predefined columns and rows. In the 1700s they were used to control textile looms and other mechanical devices such as fairground organs. In the late 1800s, Herman Hollerith developed a card system to process data from the American census. He describes the design of the cards

9. Renouf and Sinclair (1991) use the term "constructional tendency", but do not refer to Allén's work. Another example of multiple discovery.

(see Figure 3), the keypunch, and the electrical counting machine, which could handle over 100 million punch cards (Hollerith 1894, 1895, 1898).[10]

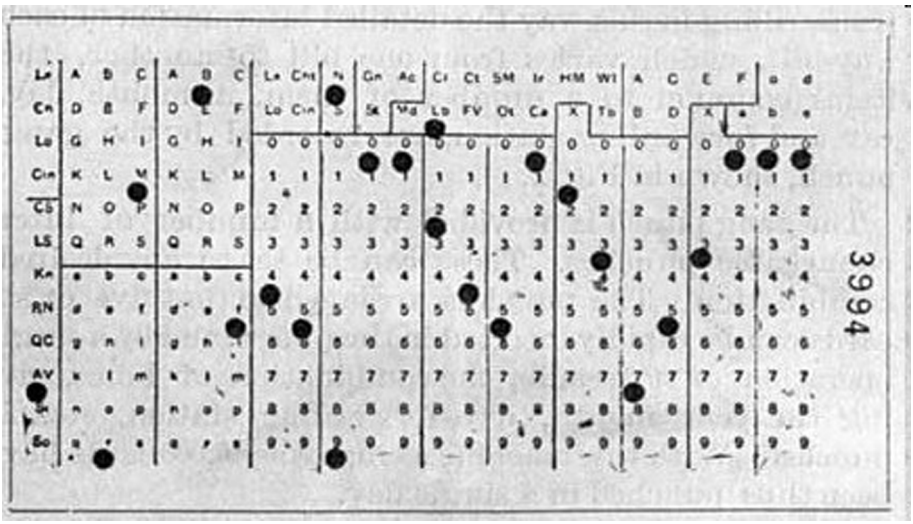

Figure 3. A Hollerith punched card. From the *Railroad Gazette*, 19 April 1895
From Wikimedia Commons

The theoretical idea was the permutation index, proposed by Andrea Crestadoro in order to make library catalogues more efficient. Crestadoro (1856) proposed that a catalogue should comprise "an encyclopaedic concordance to the title pages" of books. By permuting the words in titles, the subject index would follow the author's own definition of the contents of the book. However, Crestadoro lacked a realistic way of implementing his system.[11] This had to wait for Hans Peter Luhn, who combined the technology of punch cards with the concept of permutation, in order to construct the first automatic method of generating a concordance in what is nowadays the conventional layout, which centres node words on the page or screen.[12]

Luhn was born in Germany, worked in printing and textiles, and patented many inventions including a folding raincoat and a petrol pump. He moved to the USA, where he joined IBM in 1941, worked on information retrieval, and invented

10. His Tabulating Machine Company was one of the companies which later became IBM.

11. Crestadoro (1864) is often credited with implementing KWIC indexing based on the phraseology of the title-page. But the brief introduction describes only very sketchily "a concordance of titles" which would follow "each author's definition of his [own] book". I have to admit that I cannot see the implementation of the indexing system in the Catalogue itself.

12. For material on Luhn, I am grateful to John Sinclair, Bill Fletcher and Jutta Steckeweh.

an automatic method of indexing the titles of books and articles. The method generated a display which allowed users to see "at a glance" the relations between titles. Stop words were removed from titles, which were then rotated in a cycle, to align each content word in turn. Suppose we have the title *Introduction (to) photosynthesis (and its) applications*. We remove the bracketed words, and align the content words in turn with the same content words in other titles:

	introduction	photosynthesis applications
	introduction	experimental marine biology
introduction	photosynthesis	applications
using	photosynthesis	school biology
introduction photosynthesis	applications	
	applications	systems biology
introduction	experimental	marine biology
statistical methods	experimental	psychology

He developed this idea in two ways. He pointed out that the technique could identify the topic of an article via words which were frequent in absolute or relative terms:

> Keyword-in-context indexing […] may be applied to the title of an article, its abstract or its entire text. […] By making the keywords assume a fixed position within the extracted portions […] the KWIC index is generated.
>
> (Luhn 1960: 289)

This is close to the concept of "keywords" as used by the *WordSmith Tools* software (Scott & Tribble 2006). He also proposed a method of using the relative frequency of words to determine, entirely automatically, "which sentences of a [technical] article may best serve as the auto-abstract" (Luhn 1958). He uses mid-frequency words as the best indicators of topic, arguing that very common and very rare terms are weaker discriminators. One of his word-frequency diagrams shows that he is using Zipf's Law, though Zipf (1949) is not referenced. His overall aims are to save human effort in summarising relevant information in the increasing flood of scientific publications, and to achieve objectivity in what is otherwise an inconsistent human process. His article ends optimistically:

> If machines can perform satisfactorily […], a substantial and worthwhile saving in human effort will have been realized. The auto-abstract is perhaps the first example of a machine-generated equivalent of a completely intellectual task in the field of literature evaluation. (Luhn 1958: 165)

Fischer (1966) documents work on KWIC indexing in the 1950s and 1960s, including work apparently done by the CIA in 1952, but then kept secret till the early 1960s. Stevens (1965), Soy (1998) and Losee (2001) give further details of Luhn's work.[13]

Up until the 1970s, both data and programs were usually input into computers on standard IBM punch cards which had 80 columns and 12 rows (see Figure 4). Many early programming languages, such as FORTRAN, required specific information in specific columns of the card. For example, labels (such as GOTO or READ) had to be in columns 1 to 5, a non-blank character in column 6 was interpreted as a continuation of the previous card, statements were in columns 7 to 72, and any information in columns 73 to 80 was ignored and could be used for comments.

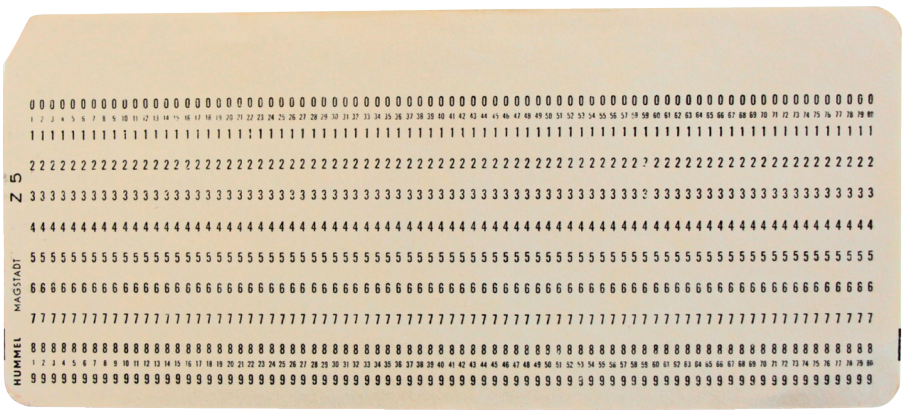

Figure 4. IBM punch card, 80 columns, 12 rows
Photograph M. Stubbs

12. Concordance packages and programming languages

From this fixed column format, it was a short step to making linguistic patterns visible by extracting words from texts and aligning them vertically on the page/screen in a few words of co-text to left and right. It is not absolutely certain when the importance of KWIC concordances was recognised by linguists, but concordance packages were available from the mid-1960s. COCOA (Count and Concordance Generation on Atlas), written in FORTRAN, was first available in

13. Luhn was long neglected by corpus linguists, but is now increasingly cited, though not by Fraser (1996), Kennedy (1998), Hockey (2004), or Barnbrook, Mason and Krishnamurthy (2013).

1967. CLOC (CoLOCation), written in ALGOL, was commissioned by Sinclair in the 1970s (Reed 1977). In the *OSTI Report* (Sinclair, Jones & Daley 1970[2004]), concordance lines have the keyword in italics (in the 2004 version), but they are not quite aligned in KWIC format.[14]

From 1981 the mainframe version of the Oxford Concordance Program (designed by Susan Hockey and Lou Burnard) was available, and from 1987 its desktop PC version, Micro-OCP, which was one of the first and most important programs for individual researchers. It processed data in batch mode, and was run from command lines, but used menus and therefore required no programming skills. It could make word lists, indexes and concordances, and a particular strength for humanities computing was its ability to handle text in different alphabets and languages by using several characters to define one letter. (On early linguistic computing, see also Berry-Rogghe and Crawford (1973), Reed (1978), Hockey (1987).)

In the 1960s Griswold, Poage and Polonsky (1968) developed SNOBOL (String Oriented Symbolic Language), the first high-level programming language specifically designed to accept text as input: simply strings of alphabetic characters. It was well described by Burnard (1978–79), Hockey (1985) and Butler (1985), and widely used in the humanities in the 1970s and 1980s, since it was extremely easy to use and very powerful: linguistic patterns can be defined as strings of characters, saved, concatenated, and used recursively within other patterns. From the late 1970s to the mid-1990s came other high level programming languages (e.g. C++, AWK, Perl, Java); from the late 1980s onwards came corpus analysis packages (e.g. *Micro Concord, TACT, WordCruncher, MonoConc, Longman Mini-Concordancer, WordSmith Tools*); and in the late 1990s came user-interfaces to specific large corpora.

It was the ability to search and redisplay large quantities of data in different formats which led to new discoveries in semantics. Modern software can represent data in many different ways, including lists (e.g. frequency, reverse alphabetical, etc.) and graphical displays (e.g. Zipf-type distributions, dispersal plots, collocate "clouds"). But the basic way of making phraseology visible is still a KWIC concordance, in which node words are aligned on the vertical axis.

A single horizontal (syntagmatic) concordance line is a fragment of *parole*, which has been artificially removed from its original natural communicative surroundings in a unique text. The vertical (paradigmatic) axis shows recurrent words which have been aligned by the linguist, and therefore shows repeated patterns across multiple texts. This formal repetition provides evidence both of phraseological units in *langue*, and recurrent collocates and/or co-occurring

14. CLOC is still available for download at <http://www.textworld.com/> (17 September 2018).

paraphrases provide observable evidence of the meaning of the units (cf. Tognini-Bonelli 2001: 3).

Once these visualisations are available, the patterns are often obvious, and verbal explanation may be of secondary importance. Cruden (1737) understood how analytic tools can play a hermeneutic role by displaying text in unfamiliar ways. And as Sinclair (1991: 100) famously said: "The language looks rather different when you look at a lot of it at once".

13. Conclusion

Over a long period of time there has been constant interaction between practice, technology and theory.

Since the 1980s, the most obvious breakthroughs in corpus analysis have been due to the ability to apply quantitative methods to large corpora. The use of statistical methods has shown that patterns of lexical co-occurrence are far more systematic than was previously realised. However, it is disputed whether statistical methods can handle the peculiar numerical properties of language data, or whether increasingly sophisticated statistics simply move further and further away from the original raw text. A question for the future is to design appropriate automatic methods for visualising language data.

But this article is about the past, so it stops here.

Acknowledgements

For critical comments on an earlier draft I am very grateful to Andreas Gestrich, Sebastian Hoffmann, Kate Tranter, and two anonymous referees.

Dates of some of the major scholars discussed

c. 265–340	Eusebius of Caesarea
1699–1770	Alexander Cruden
1709–1784	Samuel Johnson
1807–1886	Richard Chenevix Trench
1808–1879	Andrea Crestadoro
1837–1915	James Augustus Henry Murray
1843–1928	Friedrich Wilhelm Kaeding
1860–1929	Herman Hollerith
1860–1943	Jens Otto Harry Jespersen

1865–1947	Charles Bally
1874–1949	Edward Lee Thorndike
1877–1949	Harold Edward Palmer
1887–1967	Charles Carpenter Fries
1888–1973	Michael Philip West
1889–1951	Ludwig Josef Johann Wittgenstein
1890–1960	John Rupert Firth
1894–1978	Warren Weaver
1895–1961	Walter Porzig
1896–1964	Hans Peter Luhn
1900–1972	Georges Gougenheim
1910–2002	Winthrop Nelson Francis
1911–1960	John Langshaw Austin
1913–2011	Roberto Busa
1914–2005	Angus McIntosh
1920–2017	Charles Randolph Quirk
b. 1928	Sture Allén
b. 1931	Jan Svartvik
1933–2007	John McHardy Sinclair
1936–2014	Geoffrey Neil Leech
1939–2010	Stig Johansson

References

Allén, S. et al. 1975. *Nusvensk frekvensordbok baserad på tidningstext: Frequency dictionary of present-day Swedish, based on newspaper material. Vol 3. Ordfoerbindelser. Collocations*. Stockholm: Almqvist & Wiksell.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford: Clarendon Press.

Ayscough, S. 1790. *An Index to the Remarkable Passages and Words Made Use of by Shakespeare; Calculated to Point out the Different Meanings to which the Words are Applied*. London: Stockdale.

Bally, C. 1909. *Traité de stylistique française*. Heidelberg: C. Winter.

Barnbrook, G., Mason, O. & Krishnamurthy, R. 2013. *Collocation: Applications and Implications*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9781137297242

Berry-Rogghe, G. L. M. & Crawford, T. D. 1973. *COCOA: A Word Count and Concordance Generator*. Chilton: Atlas Computer Laboratory.

Brewer, C. 2011. *Examining the OED*. <http://oed.hertford.ox.ac.uk/main/index.html> (17 September 2018).

Burnard, L. 1978–1979. SNOBOL: The language for literary computing. *ALLC Journal* 6–7.
Busa, R. 1974. *Index Thomisticus: Sancti Thomae Aquinatis operum omnium indices et concordantiae …* Stuttgart-Bad Cannstatt: Frommann-Holzboog.
Busa, R. 1976. Guest editorial: Why can a computer do so little? *Bulletin of the Association for Literary and Linguistic Computing* 4(1): 1–3.
Busa, R. 1992. Half a century of literary computing: Towards a "New" philology. *Historical Social Research/Historische Sozialforschung* 17, 2(62): 124–33.
Busa, R. (ed.). 1992. *Thomae Aquinatis Opera Omnia cum Hypertextibus in CD-ROM*. Milano: Editoria Elettronica Editel.
Busa, R. 2004. Foreword: perspectives in the digital humanities. In *A Companion to Digital Humanities*, S. Schreibman, R. G. Siemens & J. Unsworth (eds), xvi–xxi. Oxford: Blackwell. <http://www.digitalhumanities.org/companion/> (10 September 2018).
Butler, C. S. 1985. *Computers in Linguistics*. Oxford: Blackwell.
Crestadoro, A. 1856. *The Art of Making Catalogues of Libraries: Or, a Method to Obtain in a Short Time a Most Perfect, Complete, and Satisfactory Printed Catalog of the British Museum Library by a Reader Therein*. London: Literary, Scientific & Artistic Reference Office. <http://books.google.co.uk/books> (1 November 2013).
Crestadoro, A. 1864. *Catalogue of the Books in the Manchester Free Library*. Manchester Public Libraries (Manchester, England). London: Sampson, Low, Son, & Marston. <http://books.google.co.uk/books> (10 September 2018).
Crowley, T. 1989. *The Politics of Discourse: The Standard Language Question in British Cultural Debates*. Houndmills: Macmillan. https://doi.org/10.1007/978-1-349-19958-7
Cruden, A. 1737. *A Complete Concordance to the Holy Scriptures of the Old and New Testament; Or a Dictionary and Alphabetical Index to the Bible …* London: Frederick Warne & Co.
Cruden, A. 1741. *A Verbal Index to Milton's Paradise Lost. Adapted to Every Edition but the First, Which was Publish'd in Ten Books Only*. London: W. Innys & D. Browne.
Eusebius. c. 320?. *Epistula ad Carpianum ad canones evangeliorum praemissa*. Greek with English translation.<http://en.wikipedia.org/wiki/Epistula_ad_Carpianum> (17 September 2018).
Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, 1-32. Philological Society.
Fischer, M. 1966. The KWIC index concept: A retrospective view. *American Documentation*, April, 57–70. https://doi.org/10.1002/asi.5090170203
Francis, W. N. 1992. Language corpora BC. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*, Stockholm, 4–8 August 1991, J. Svartvik (ed.). Berlin: De Gruyter. https://doi.org/10.1515/9783110867275.17
Francis, W. N. & Kučera, H. 1964. *Brown Corpus Manual. Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence RI: Brown University. <http://clu.uni.no/icame/manuals> (17 September 2017).
Fraser, M. 1996. *A Hypertextual History of Humanities Computing: The Pioneers*. <http://users.ox.ac.uk/~ctitext2/history/pioneer.html> (17 September 2018).
Fries, C. C. 1952. *The Structure of English*. New York NY: Harcourt Brace.
Fries, P. H. 2012. Charles C. Fries, linguistics and corpus linguistics. *ICAME Journal* 34: 89–119.
Giegerich, H. 2005. *Obituary for Angus McIntosh (1914–2005)*. <http://www.inf.ed.ac.uk/events/amcintosh.html> (17 September 2018).

Gougenheim, G., Michea, R., Rivenc, P. & Sauvageot, A. 1956. *L'élaboration du français élémentaire: Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier. (Revised ed. retitled *L'élaboration du français fondamental*, 1964, Paris: Didier).

Griswold, R. E., Poage, J. F. & Polonsky, I. P. 1968. *The SNOBOL4 Programming Language*. Englewood Cliffs NJ: Prentice Hall.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge MA: The MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Hockey, S. M. 1987. *Micro-OCP (Oxford Concordance Program)*. Oxford: OUP.

Hockey, S. M. 1985. *SNOBOL Programming for the Humanities*. Oxford: Clarendon Press.

Hockey, S. M. 2004. The history of humanities computing. In *A Companion to Digital Humanities*, S. Schreibman, R. G. Siemens & J. Unsworth (eds), Oxford: Blackwell. <http://www.digitalhumanities.org/companion/> (17 September 2018). https://doi.org/10.1002/9780470999875.ch1

Hollerith, H. 1894. The electrical tabulating machine. *Journal of the Royal Statistical Society* 57 (4): 678–89. https://doi.org/10.2307/2979610

Hollerith, H. 1895. Hollerith's electric tabulating machine. *Railroad Gazette*, 19 April 1895.

Hollerith, H. 1898. *Art of compiling statistics. No. 395,782*. United States Patent Office. <https://www.google.de/patents> (1 November 2013).

Howatt, A. P. R. 2004. *A History of English Language Teaching*, 2nd ed. with H.G. Widdowson. Oxford: OUP.

Hüllen, W. 1996. Schemata der Historiographie. Ein Traktat. *Beiträge zur Geschichte der Sprachwissenschaft* 6(1): 113–125. Also in M. Isermann (ed.). 2002. *Werner Hüllen: Collected Papers on the History of Linguistic Ideas*, 16–28. Münster: Nodus.

Jespersen, O. 1909[1949]. *A Modern English Grammar on Historical Principles*. Heidelberg: C. Winter.

Johansson, S. 2008. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In *Corpus Linguistics: An International Handbook*, Vol. 1, A. Lüdeling & M. Kytö (eds), 33–52. Berlin: De Gruyter.

Johnson, S. 1747. The plan of an English Dictionary, J. Lynch (ed.). <http://andromeda.rutgers.edu/~jlynch/Texts/plan.html> (17 September 2018).

Johnson, S. 1755. *A Dictionary of the English Language: In Which the Words are Deduced from Their Originals, and Illustrated in Their Different Significations by Examples from the Best Writers …* London: Knapton.

Kaeding, F. W. 1897. *Häufigkeitswörterbuch der deutschen Sprache: Festgestellt durch einen Arbeitsausschuss der deutschen Stenographie-Systeme*. Steglitz bei Berlin.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Keay, J. 2004. *Alexander the Corrector: The Tormented Genius who Unwrote the Bible*. London: HarperCollins.

Kuhn, T. S. 1969. Comment on the relations of science and art. *Comparative Studies in Society and History* 11: 403–412. https://doi.org/10.1017/S0010417500005466

Leech, G. 1991. The state of the art in corpus linguistics. In *English Corpus Linguistics*, K. Aijmer & B. Altenberg (eds), 8–30. London: Longman.

Leech, G. 2013. The development of ICAME and the Brown family of corpora. In *The Many Facets of Corpus Linguistics in Bergen: In Honour of Knut Hofland* [Bergen Language and Linguistics Studies 3. 1], L. Hareide, C. Johansson & M. Oakes (eds). <https://bells.uib.no/index.php/bells/article/view/358/373> (17 September 2018). https://doi.org/10.15845/bells.v3i1.358

Léon, J. 2005. Claimed and unclaimed sources of corpus linguistics. *Henry Sweet Society Bulletin* 44: 36–50.

Liberman, M. 2004. A brief and a compendious table. *Language Log*, 4 March 4 2004. <http://itre.cis.upenn.edu/~myl/languagelog/archives/000537.html> (17 September 2018).

Losee, R. M. 2001. Term dependence: A basis for Luhn and Zipf models. *Journal of the American Society for Information Science and Technology* 52(12): 1019–1025. https://doi.org/10.1002/asi.1155

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2): 159–165. https://doi.org/10.1147/rd.22.0159

Luhn, H. P. 1960. Keyword-in-context index for technical literature. *American Documentation* xi(4): 288–95. https://doi.org/10.1002/asi.5090110403

McIntosh, A. 1961. Patterns and ranges. *Language* 37: 325–337. https://doi.org/10.2307/411075

Meehan, B. 1994. *The Book of Kells: An Illustrated Introduction to the Manuscript in Trinity College, Dublin*. London: Thames & Hudson.

Meyer, C. F. 2008. Pre-electronic corpora. In *Corpus Linguistics: An International Handbook*, Vol. 1, A. Lüdeling & M. Kytö (eds), 1–13. Berlin: De Gruyter.

Moon, R. 2007. Sinclair, lexicography, and the Cobuild Project: The application of theory. *International Journal of Corpus Linguistics* 12(2): 159–181. https://doi.org/10.1075/ijcl.12.2.05moo

Mugglestone, L. C. 2005. *Lost for Words: The Hidden History of the Oxford English Dictionary*. New Haven CT: Yale University Press.

Murray, K. M. E. 1977. *Caught in the Web of Words: James A H Murray and the Oxford English Dictionary*. New Haven CT: Yale University Press.

Oliver, H. H. 1959. The epistle of Eusebius to Carpianus: Textual tradition and translation. *Novum Testamentum* 3(1–2): 138–145.

Palmer, H. E. 1933. *Second Interim Report on English Collocations* (submitted to the Tenth Annual Conference of English Teachers, Institute for Research in English Teaching, Dept. of Education, Tokyo). Tokyo: Kaitakusha.

Porzig, W. 1934. Wesenhafte Bedeutungsbeziehungen. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 58: 70–97.

Ramsay, S. 2008. Algorithmic criticism. In *A Companion to Digital Humanities*, S. Schreibman, R. G. Siemens & J. Unsworth (eds). Oxford: Blackwell. <http://www.digitalhumanities.org/companionDLS/> (17 September 2018).

Rastall, P. 2001. Richard Chevenix Trench: More than just a populariser. *The Henry Sweet Society Bulletin* 37: 22–39.

Reed, A. 1977. CLOC: A collocation package. *Association for Literary and Linguistic Computing Bulletin* 5(2): 168–173.

Reed, A. 1978. *CLOC User Guide*. Birmingham: University of Birmingham, Computer Centre.

Renouf, A. 2007. Corpus development 25 years on: From super-corpus to cyber-corpus. In *Corpus Linguistics 25 Years On*, R. Facchinetti (ed.), 127–149. Amsterdam: Rodopi. https://doi.org/10.1163/9789401204347_004

Renouf, A. & Sinclair, J. 1991. Collocational frameworks in English. In *English Corpus Linguistics*, K. Aijmer & B. Altenberg (eds), 128–43. London: Longman.

Scott, M. & Tribble, C. 2006. *Textual Patterns* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.22

Sellar, W. C. & Yeatman, R. J. 1931. *1066 and All That; A Memorable History of England, Comprising All the Parts You Can Remember Including One Hundred and Three Good Things, Five Bad Kings and Two Genuine Dates*. London: Methuen.

Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. McH. 2004. Interview with John Sinclair conducted by Wolfgang Teubert. In *The OSTI Report*, R. Krishnamurty (ed.), xvii–xxix. London: Continuum.

Sinclair, J. McH. (ed.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: Collins ELT.

Sinclair, J. McH., Jones, S. & Daley, R. 1970[2004]. *English Collocation Studies*. Original mimeoed report 1970. Re-published as Krishnamurthy, R. (ed.). 2004. *English Collocation Studies: The OSTI Report*. London: Continuum.

Soy, S. K. 1998. *Class notes: H. P. Luhn and automatic indexing*. <https://www.ischool.utexas.edu/~ssoy/organizing/luhn.htm> (17 September 2018).

Sperberg-McQueen, C. M. & Burnard, L. (eds). 1990. *Guidelines for the Encoding and Interchange of Machine-Readable Texts*. TEI P1. Draft 1.1. Chicago-Oxford. Updates: <http://www.tei-c.org/Vault/ED/edp01.xml> (17 September 2018)

Stevens, M. E. 1965. *Automatic Indexing: A State of the Art Report*. <http://sigir.org/files/museum/monograph-91/pdfs/p1-part_1.pdf> (17 September 2018). https://doi.org/10.6028/NBS.MONO.91

Svartvik, J. 2007. Corpus linguistics 25+ years on. In *Corpus Linguistics 25 Years On*, R. Facchinetti (ed.), 11–25. Amsterdam: Rodopi. https://doi.org/10.1163/9789401204347_003

Teubert, W. 2004. A brief history of corpus linguistics. In *Lexicology and Corpus Linguistics*, M. A. K. Halliday, W. Teubert, C. Yallop & A. Čermáková, 107–112. London: Continuum.

The Oxford English Dictionary. 1933. James A. H. Murray (ed.). Oxford: Clarendon Press.

Thorndike, E. L. & Irving, Lorge. 1944. *The Teacher's Word Book of 30,000 Words*. New York NY: Teachers College, Columbia University.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Trench, R. C. 1857. *On Some Deficiencies in our English Dictionaries: Being the Substance of Two Papers Read Before the Philological Society, Nov. 5, and Nov. 19, 1857*. Philological Society (Great Britain). London: J.W. Parker & Son.

Voloshinov, V. N. 1929[1973]. *Marxism and the Philosophy of Language*, transl. by L. Matejka & I. R. Titunik, first published in Russian 1929. New York NY: Seminar Press.

Weaver, W. 1955. Translation. In *Machine Translation of Languages*, W. N. Locke & D. A. Booth (eds), 15–23. Cambridge MA: The MIT Press. 1949 version at: <http://www.mt-archive.info/Weaver-1949.pdf> (17 September 2018).

West, M. 1953. *A General Service List of English Words*. London: Longman, Green & Co.

Winchester, S. 1999. *The Surgeon of Crowthorne: A Tale of Murder, Madness and the Oxford English Dictionary*. London: Penguin.

Winchester, S. 2003. *The Meaning of Everything: The Story of the Oxford English Dictionary*. Oxford: OUP.

Wisbey, R. A. (ed.). 1971. *The Computer in Literary and Linguistic Research*. Cambridge: CUP.

Wittgenstein, L. 1953. *Philosophische Untersuchungen*. Frankfurt: Suhrkamp.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Reading MA: Addison-Wesley.

Modes of analysis

An autoergography of corpus linguistics from lexis to discourse

Ramesh Krishnamurthy
Aston University

Throughout history, the study of language has involved looking for units, their patterns of combination, the functions they serve, and the processes involved; and construing the relationships between these features and the meanings which arise from them. Developments in technology inevitably revolutionise ideas, innovate methods, and transform the fields in which they are implemented. So, just as the invention and development of writing systems, and the advent of printing, did in the distant past, computers have had a similarly revolutionary effect on linguistics in recent decades. This paper looks at some of the ways in which corpus linguistics has used the latest technologies to embark on a substantial re-investigation and re-appraisal of the elements of language, their roles within the language system, and their relationships within the wider social context of human beings, their environments, and their activities. The structuring of the paper owes much to recent retrospection, but the pieces of research consist mainly of my early explorations of corpus linguistics, hence '*auto* (self) + *ergo* (work) + *graphy* (description)' in the title.

1. Introduction

Whether we consider spoken or written language, there are many postulated units to choose between. The most widely accepted unit across languages and time seems to be 'word', and more recently 'discourse' is gaining cross-disciplinary acceptance. Linguists also discuss 'phonemes', 'syllables', 'tone units', 'prosodic/intonation units', 'utterances' in speech, 'characters/letters/graphemes', 'clauses', 'sentences', 'paragraphs', 'sections', 'chapters', and 'texts' in writing (as well as 'lines' and 'verses/stanzas' in poetry, and other specialist terms in other genres). Various forms of speech may also be measured in units of time (e.g. a *three-minute* presentation, a *one-hour* lecture). Units are selected for their ease of identification (e.g. 'words' in

https://doi.org/10.1075/scl.87.03kri

written English are separated by spaces or punctuation marks), but often involve subjective judgement (e.g. 'morphemes', 'multi-word units'), or some agreed convention or consensus (e.g. between academics, lexicographers, and grammarians; proofreaders and publishers; in-house and popular style guides).

Linguists frequently discuss 'patterns', e.g. Lukin (2014) quotes Firth as saying that linguistics involved "patterns of life", Miller (1956) says we organise units of information into patterns to remember larger chunks, Sinclair (2004: 19) discusses "pattern of morphemes" and "collocational patterns" (Sinclair, Jones & Daley 1970[2004]: xix), and Francis, Hunston and Manning (1996) introduce 'pattern grammar'. Teubert (2010: 8–9) suggests that late 19th century linguistics tried to gain scientific credence "by assimilating the discovery of linguistic patterns to the discovery of laws of Nature" and sacrificing "any distinction between rules and regularities (Harris 1987: 109)". Functions are highlighted, e.g. by Halliday's Systemic Functional Grammar, Sinclair's (Sinclair, Jones & Daley 1970[2004]: xxi) discussion of the selective and focusing functions of adjectives, and assertion that semantic prosody shows how a piece of language is "to be interpreted functionally. Without it, the string of words just 'means'" (Sinclair 1996, 2004: 34). Teubert (2010: 122) discusses "what functions and purposes can be ascribed to texts". Processes are an equally pervasive linguistic focus, e.g. Firth (1957b) discusses meaning in terms of "the verbal process", "the word process", and "phonematic and prosodic processes", Sinclair (1991: 8) situates collocation in the process of meaning-creation.

Pinpointing the relationship between language units and meaning is highly problematic. Sinclair discusses collocation as "an important part of the patterning of meaning", he shifts from "the notion of the word as the unit of meaning" to "a phrasal unit", and he realizes that "*dark night* has its own meaning". He asserts that "numbers are not sensitive to meaning", that "grammar is not involved in the creation of meaning, but… the management of meaning", that "[m]eaning is an impression in the mind of an individual, and that is impenetrable, using linguistic techniques … entirely provisional … built up of our interpretations", which are "neither fully predictable nor formalisable" and that the task of linguists lies in "fitting the forms to the meanings, and not the other way round" (Sinclair, Jones & Daley 1970[2004]: xvii–xxix). Teubert (2010: 2) concurs:

> The word *life* means what life is for us. The meaning of life is therefore not really different from the meaning of the word *life*. It is all that has been said about it… It is difficult to imagine that by pondering we would find an answer to the meaning of life that is not already expressed in the discourse.

Corpus linguistics has supplemented traditional, manual approaches with computational ones, added discrete observational analysis to subjective and intuitive judgements, enhanced the impressionistic with the statistical, and improved

evaluation by adding precision and recall to human satisfaction. It has also initiated the task of defining new units, and continued the process of redefining traditional linguistic units, e.g. "[a]ny instance of language that is operational, as distinct from citational … is text. The term covers both speech and writing" (Halliday 2006: 284). After a brief look into a few relevant aspects of the history of linguistics, this paper will focus on some of the units that corpus linguistics has analysed, linking the recent work back to history where appropriate.

2. Ancient and modern linguistics

Holden (2004: 1316) estimates that humans evolved their speech-enabling anatomy c. 300,000 years ago, the speech "gene" c. 200–100,000 years ago, and "fully developed language" c. 50,000 years ago. Atkinson (2011) uses phonemic diversity to add linguistic support to the prevalent view among geneticists and anthropologists that suggests an African origin for modern humans. The earliest ideas about language must therefore have occurred within oral cultures. Writing systems were not developed until much later (c. 3500 BC), and most of the earliest written evidence has perished or been destroyed. Hence, many of our ideas about ancient linguistics can only be inferred or conjectured from fragmentary texts, or reconstructed from much later sources.

Teubert, perhaps because of his focus on discourse (see Section 11), suggests that in oral societies, "language is a social practice that people are mostly unaware of" (Teubert 2007: 3) because in them, "people do not ask: 'What does this (piece of) text mean?' but 'What do you mean?'" (Teubert 2003: 134). Thus "the hearer is less interested in what a text means than in what the speaker wants to convey" (Teubert 2003: 144). Spoken language can only "refer (and react) to the preceding and accompanying contributions to the discourse and also to other, more or less symbolic forms of social interaction" (Teubert 2003: 137). Speech also "organises the activities in which the group members engage, by means of giving instructions, by asking for and providing information, and by telling people what to do" and "establishes, preserves or seeks to alter the balance among the members of the discourse community" (Teubert 2003: 137). He concludes "Even when we acknowledge that speech is the primordial form of language, it seems we cannot talk about it without taking refuge in the changes that writing brought about" (Teubert 2003: 137).

Schmandt-Besserat (1977) says that the Sumerians by the end of the 4th millennium B.C. had "developed a system of recording numerals, pictographs and ideographs on specially prepared clay surfaces", a pictograph being "realistic" and an ideograph being an "abstract" symbol. These had been preceded from the early

8th millennium B.C. by counting tokens, notched or marked to denote their value, of which Schmandt-Besserat (1991) says "each type of token gave rise to a different type of sign in the Sumerian script and can be considered as a separate precursor of writing". Numerals were therefore the earliest symbolic characters in writing. Teubert (2003: 135) suggests a similar process: "the signs of the early writing systems … refer to real world objects". The systems were language-independent, so the characters did not represent spoken sounds. Gradually, language evolved from "a transitory acoustic phenomenon" into "a material thing, into clay tablets. It became something that could be touched, stored, re-awakened to speech and negotiated" (Teubert 2003: 135). However, Teubert adds, "with the evolution of more sophisticated methods of arranging, varying and permuting the characters, writing could be made to resemble speech" (Teubert 2003: 136). The technology for recording sound would not be invented until the late 19th century A.D., so humans had to record sounds as graphic symbols instead.

The Chinese had evolved their own stable ideographic writing system by c. 1500 B.C. (Robins 2013: 122), but did not make the attempt to link their characters with speech sounds, so Chinese linguists focussed on the characters (Wang 2008: 504) and the writing system (Li 2010: 8) for centuries. The first work to focus on meanings rather than forms, the *Erya* was created c. 202 B.C. to 9 A.D. (Li 2010: 121). A resurgence of interest in phonology occurred from c. 200 A.D. (Zhu 2013) due to the introduction of Indian linguistics by Buddhist monks (Robins 2013: 12).

Writing was adopted even later in the Indo-European languages, perhaps c. 900 B.C. for Greek, but the Greeks focussed mainly on meaning rather than forms. Sanskrit was not written down for many further centuries, and perhaps that is why we can see the transition from speech to writing more clearly in Sanskrit: the syllabary is ordered by the place of articulation of consonantal sounds and the system prioritises vowel-gradation (*guṇa, vṛddhi*) and vowel-conjunction (*sandhi*); the written forms merely re-articulate the speech sounds. Müller (1860) claimed that only the Hindus and Greeks developed independent systems of logic and grammar in the ancient world, and that

> [t]he Greeks began with philosophy, and endeavoured to transfer their philosophical terminology to the facts of language. The Hindus began with the facts of language, and their generalisations never went beyond the external forms of speech. Thus the Hindus excel in accuracy, the Greeks in grasp.
>
> (Müller 1860: 158–9)

Matilal (1990: 7) suggests that ancient Indian science and philosophy initially focussed less on mathematics, prioritising grammar or linguistics instead. Basham (1967: 389–90) says Indian linguists had recognised verbal roots as the primary

linguistic elements and "classified some 2,000 monosyllabic roots … thought to provide all the words of the language", and indeed Yāska's Nirukta (c. 500 B.C) lists words, their origins, meanings, and categories. However, Yāska saw the relationship between word and meaning as purely conventional and contextual (sentential), whereas others thought that a word in isolation had an inherent and eternal meaning (Matilal 1990: 106–7, Basham 1967: 392). For example, Pāṇini (c. 4th century B.C.) links semantics to syntax (thematic role of verbal arguments) and morphology (with its associated phonology), and his grammar attributes meanings to morphemic forms. Matilal (1990: 106–7) also notes that the Sanskrit word for 'grammar' (*vyākaraṇa*) has a broader, literal meaning of 'analysis'. This early Sanskrit evidence conflicts with Teubert's assertion that "[o]ral societies do not have institutions in which they learn about language" (Teubert 2003: 137), as some form of linguistic education must have existed in ancient India, although writing was not adopted in India until much later.

Plato (428/423–348/347 B.C.), unlike Yāska, suggests that "everything has a right name of its own, which comes by nature … a name is not whatever people call a thing by agreement … there is a kind of inherent correctness in names, which is the same for all men" (Fowler 1921: 383a). Aristotle (384–322 B.C.) opposes Plato in saying that "meanings are determined by habit and convention" (McKeon 1946: 195) and in giving primacy to material over immaterial entities, and distinguishes simple linguistic entities (*horse, man, fights*) from those involving composition and structure (*a man fights, the horse runs*). The Romans mainly followed the Greek traditions, and Greek grammars and dictionaries were succeeded by Latin ones. Basham (1967) concludes his description of the morphological and phonological complexities of ancient Greek and Sanskrit with the generic observation that "[a]s long as it is spoken and written a language tends to develop, and its development is generally in the direction of simplicity" (Basham 1967: 391). This is certainly evident in modern Indo-European languages in both India and Europe.

Teubert attaches a much profounder significance to the advent of writing: "The invention of writing extended the functions of language far beyond the possibilities of speech in oral societies. Language today is more than speech was then" (Teubert 2003: 142). Writing enabled us to "speculate about the relationship between the sign and the thing" and made us "aware of language" (Teubert 2003: 137). Indeed, writing "brought the concept of meaning into the language. Only literal societies can reflect on what is being said" (Teubert 2003: 136). Unlike oral societies, "[i]n the absence of a speaker who can be questioned, the precise meaning of a text has to be established without recourse to the speaker's intentions" (Teubert 2003: 144), "the text has become dislodged from its author" (Teubert 2003: 134) and "the existence of written texts … enabled us to reflect about our linguistic

practice" (Teubert 2003: 138). Language has become "a store of socially negotiated meaning" (Teubert 2003: 138) and "a discourse object, a topic one could talk about" (Teubert 2007: 3).

As Christianity spread through Europe, monks copying religious texts added marginal notes and translations ('glosses') in their own vernacular languages. Secular texts were later treated similarly, and the various glosses were collected and listed, grouped thematically initially, then alphabetically (in dictionaries). Hence the focus was on meanings rather than forms: the aim was encyclopaedic not linguistic and the vernaculars lacked standardised spelling systems. The adoption of the Latin alphabet by many modern European languages has created various problems, e.g. of divergent pronunciations for the same letters (e.g. English, Italian, German, Spanish, French), of widespread need for diacritics (e.g. Polish, Czech), or of loss of letters (e.g. *thorn* (Weisser 2016: 15), *eth, yogh, ash*, etc. from Old English). Modern English uses 26 letters to represent 44 sounds (unlike Sanskrit, whose sounds and characters are an exact match). As dictionaries became commonplace and spelling standardised, meaning was notionally vested in the word, initially merely as a lexicographic indexical convention, but took root axiomatically in many scholarly works as well. The word – in isolation – also became the focus of most phonetic and phonological studies. Only fairly recently has linguistics started investigating spoken units larger than the word, such as tone/intonation units and utterances.

Many ancient linguistic questions – e.g. which units to study; are the units componential; what are the links between language, thought and meaning; what are the processes involved – are still being discussed by more recent linguists. Humboldt (1767–1835) on the one hand asserts that writing was "the material embodiment of the specific formative principle a language employs to construct meaning", but on the other hand that "language in actuality only exists in spoken discourse, its grammar and dictionary are hardly even comparable to its dead skeleton" (SEP 2011). Chomsky (1968, transcribed by Blunden 1998) justifies his own approach by saying that it "provides an explication of the Humboldtian idea of 'form of language'", citing Humboldt's definition of language as a system "where the laws of generation are fixed and invariant, but the scope and the specific manner in which they are applied remain entirely unspecified".

Bloomfield (1887–1949) excluded semantics from linguistics: "The study of language can be conducted without special assumptions so long as we pay no attention to the meaning of what is spoken" (Bloomfield 1933: 75). He also excludes words: the lexicon is merely "an appendix of grammar and the list of basic irregularities" (Bloomfield 1933: 274). Robins (1964: 18–19) agreed: "The categories of phonetics, phonology and grammar are general; the components of the lexicon of a language are particular". Atkins and Rundell (2008: 49) view this

as a philosophical dichotomy: "Lexicographers (and corpus linguists generally) are empiricists" who describe linguistic performance by observing usage; rationalists, on the other hand, aim to describe linguistic competence, "the internalized, but subconscious, knowledge we have of the rules" of a language, obtained by introspection. Church (2011: 2) sees this dichotomy as a historical "oscillation between Rationalism and Empiricism" (the 1950s and 1990s empirically-focussed, the 1970s and 2010s rationalist-oriented).

3. Corpus linguistics

London University became central in the development of the usage-based view of language in the 1930s. The anthropologist Malinowski emphasised the role of context in interpreting social behaviour, including language: "a word like this can … be understood only in the context of its phrase … its fuller meaning becomes intelligible only in the context of the native life and of native sociology" (Malinowski 1922: 458–9). And his linguist colleague Firth added that "'context of situation' is best used as a suitable schematic construct to apply to language events" (Firth 1957b, cited in Louw 2000). Firth's schematic construct included the participants, their verbal and non-verbal actions, any objects involved, and the effect of the actions. Godart-Wendling (2014) asserts that Firth combines ideas from Wittgenstein (meaning as use, language acquisition, language as a set of games) with ideas from Malinowski (context of situation, language as a mode of action).

In *Modes of Meaning* (the inspiration for the current paper), Firth (1957b) discusses the various linguistic levels at which meaning arises, starting with "the verbal process in the context of situation", and includes syntax ("word process in the sentence") and phonology ("phonematic and prosodic processes within word and sentence") as well as orthographic, phonetic, grammatical, formal, and etymological levels. De Beaugrande (1991) says Firth's later works mention twenty-four "levels" (sometimes named "modes"). Sinclair (1966: 410–11) asserts "lexis as an independent part of language form", which "describes the tendencies of items to collocate with each other" and "cannot be described in terms of small sets of choices" (unlike grammar, "a large set of choices … from a small list of possibilities"), and recognises the need for computers in its study. Halliday (1966: 148) also stresses the importance of lexis, adducing Firth in support. Firth (1957a: 190) further states that "the main aim of descriptive linguistics is to make statements of meaning", and Willis (2003: 16) says "[f]or Halliday the important thing about language is the capacity to mean". Some other linguists broadly agree: words are "the pre-set meaning-bearing units of a language" (Bolinger 1968: 105),

and "[w]hile without grammar little can be conveyed, without vocabulary nothing can be conveyed" (Wilkins 1972: 111–2).

By the 1960s, the main elements of corpus linguistics had coalesced: the empirical study of language in use, the importance of words, and the focus on meaning. The final requisite, computers, started to become available around the same time, as a result of codebreaking techniques and technological developments during World War Two. The first phase of corpus linguistics began at London University (*Survey of English Usage*), Brown University (*Standard Corpus of Present-Day American English*), and Edinburgh University (John Sinclair's work on a spoken corpus, focussing on lexis and collocation).

Sinclair moved to Birmingham University in 1965, continued to work with corpora, and introduced corpus linguistics to his colleagues and students. In 1980, he initiated the COBUILD (*Collins Birmingham University International Language Database*) project, funded by Collins the publishers, and embarked on a pioneering and comprehensive corpus-driven description of English, made publicly available in several dictionaries, grammars, usage books, and teaching and research publications. It was the largest commercially-funded humanities project at any UK university, and it published the first dictionary in history to be based on corpus analysis. COBUILD's innovative corpus methodology has been adopted by most ELT publishers, as well as many bilingual and native-speaker dictionaries, especially national dictionaries (e.g. OED). COBUILD created the largest general corpus of its time (the *Bank of English*), as well as specialised corpora (e.g. of business, academic, and pedagogic texts) and corpora in several other languages, before the project closed in 2003. Krishnamurthy (e.g. 1987, 1992b, 2000, 2002b) has written extensively about its methodology, corpora, and software. Birmingham University (2014)[1] listed COBUILD among the ten research projects in the past century that have had the greatest global impact.

4. Character mode

Unlike the Chinese, English linguists have paid a great deal of attention to phonemes, phonology and phonetics, but much less to orthographic characters, as the letters are deemed only to represent sounds (despite their complex relationships), and to make only a minimal contribution to meanings.

1. Birmingham University (2014) – *University of Birmingham's "top ten" celebrates global impact*. <http://www.birmingham.ac.uk/news/latest/2014/05/University-of-Birminghams-top-ten-celebrates-global-impact.aspx> (1 July 2017).

By 1992, COBUILD was exploring new possibilities for online interaction with its rapidly growing corpus, so I conducted a character-level analysis of the 18m-word ('m' will be used for 'million' throughout this paper in relation to corpus statistics) *Birmingham Collection of English Text* (BCET). All word types were lower-cased in frequency lists (sacrificing the – far less frequent – proper nouns in order to eliminate the vast number of words whose initial letters had been capitalised only because they occurred at the beginning of a sentence), so the list consisted of 29 single-character items (*a-z*, space, tab, and newline). The list of adjacent character-pairs contained 669 items (i.e. combinations of the 29 single-character items), and the 3-character list had 29,479 items.

Before more detailed study, the spoken data was excluded (as the transcription conventions were devised by COBUILD purely for lexicographic purposes). The 1-character frequency list for the written data (16.78m-words) was: [space] *e t a o i n s r h l d u c m* [newline] *f w g y p b v k x j q z* [tab]. The top ten 2-character sequences were: *th*, *he*, *in*, *er*, *an*, *re*, *on*, *en*, *at, nd*. Some of these items may represent grammatical words (*an, in, on, at, he*), but most are parts of longer words (e.g. ***an**tique*, *comp**any***, *beg**an***). The top ten 3-character sequences (_ represents 'space') were: *_th*, *the*, *he_*, *nd_*, *_an*, *and*, *ed_*, *_of*, *ing*, *_to*. Similar character sequences (with their pronunciations) were extracted from previous COBUILD publications and used in the *COBUILD English Guides 8: Spelling* (Payne 1995), which covered the 25,000 most frequent words in the 323m-word *Bank of English*.

Corpora have also been used in sub-lexical semantic research, i.e. phonaesthetics (e.g. Otis & Sagi 2008, Abramova, Fernández & Sangati 2013). "Phonaesthemes (Wallis 1699, Firth 1930) are frequent sound-meaning pairings, which cannot be defined entirely in terms of contrast" (Bergen 2004: 290). Bergen (2004) uses the *Brown corpus* frequency list to determine that 39% of types and 60% of tokens beginning in *gl-* have meanings related to *light* or *vision* in *Webster's Dictionary*.

5. Morpheme mode

Sinclair clarifies the reason for the gap at morphological level in the English corpus linguistics tradition. Although he acknowledges the morpheme ("Every morpheme in a text must be described" (Sinclair 1966: 423)), he says it has no empirical validity, as it is merely "one part of an orthographic word divided conventionally, e.g. *go/ing*" (Sinclair, Jones & Daley 1970 [2004]: 9), and "even the bedrock assumptions of linguistics, like the … assignment of morphological division… are not at all standardized" (Sinclair 1991: 21). Additionally, because like Firth he focusses on meaning, Sinclair (2004: 19) says it is "impossible to say precisely where the realization of that meaning starts and stops, or exactly which pattern of morphemes

is responsible for it". He says the focus on morphemes comes from American linguistic models, which in "the first half of this century put forward the morpheme, the smallest unit of grammar, as a more suitable foundation", so they "concentrated on the smaller units of language" (Sinclair 2004: 24–25). Sinclair's search for larger units includes 'polymorphemic' items, for which he gives *take the bull by the horns* as an example (Sinclair 1966: 420). Ultimately, he says "Linguistics usually operates with … abstract categories … But … it is good policy to defer the use of them for as long as possible, to refrain from imposing analytical categories from the outside" (Sinclair 1991: 29). This position is confirmed in the *COBUILD Dictionary* (Sinclair 1987a), which does not have entries for *morpheme* or *morphology*, nor are these terms used in the front matter. The terms are also absent from the Index of *Looking Up* (Sinclair 1987b), the Glossary and Index of the *COBUILD Grammar* (Sinclair 1990; *morpheme* occurs only in the Grammar Chart, xxiv–xxv), and from the *COBUILD English Guides 2: Word Formation* (Bradbury 1991).

Traditional dictionaries tend to give many of the possible morphological forms of a headword at the beginning (e.g. *encrust, ~ed, ~ing, ~s*) or end of an entry (e.g. *lame 1, 2, 3, 4, 5…; ~ly*), suggesting that they all have an equal likelihood of being used, and can be used in any of the word senses. COBUILD discovered that this was very far from the truth. Some forms are used very frequently, some are used so rarely that even very large corpora can provide little evidence for them (e.g. *encrusted* occurring 716 times, *encrusting* 90 times, *encrust* 25 times, *encrusts* 4 times in the 1.9 billion-word *Glowbe* corpus).[2] Similarly, the distribution of word-forms across word senses can be very variable, e.g. hardly any examples of *lamely* concern *injured legs*, most involve unconvincing arguments or behaviour ("*He's not finished yet." I lamely replied; Adam lamely went with his bro; a technique he lamely defends*).[3]

Krishnamurthy (1992a) illustrated the value of corpus word-frequency lists in inductive and autonomous pedagogical activities relating to morphology (see Table 1).

Table 1. BCET corpus (18m words, 1986): most frequent words ending in *–ness*

business	3,925	*awareness*	404	*madness*	222	*bitterness*	169
darkness	969	*witness*	396	*wilderness*	218	*readiness*	158
consciousness	609	*weakness*	364	*kindness*	206	*tenderness*	158
illness	608	*sickness*	292	*ness*	202	*effectiveness*	156
happiness	491	*goodness*	242	*loneliness*	200	*sadness*	153

2. <http://corpus.byu.edu/glowbe/>

3. Examples from *Glowbe* corpus. <http://corpus.byu.edu/glowbe/>

Even such minimal outputs (Table 1) can provide models for any linguistic feature, including both dominant patterns, common variations, and exceptions; e.g. after recognising *~ness* as a suffix, users can see (or be guided to notice) the adjective > abstract noun pattern (*dark* > *darkness*; *conscious* > *consciousness*; *ill* > *illness*), standard orthographic change (*happy* > *happiness*), semantic change (*business* > state of being *busy* > commercial activity), and exceptions (*witness, wilderness, ness*), which can instigate further studies using concordances, which would also be used to verify/discover the meanings of the pattern-forming items already seen/noticed.

For more highly inflected languages, morphological analysis is useful or necessary even before generating word frequency lists, so morphosyntactic analysers are often the first corpus tool developed. Other conventions in English corpus analysis may need to be revised (e.g. lowercasing all word types (see Section 6) would lose valuable grammatical information in German, which uses initial capital letters for all nouns).

The need for morphological analysis became evident when Krishnamurthy attempted to test the assertion made by several translation scholars (e.g. Toury 1980, 1995; Baker 1993) that translated works are longer than source texts because of explicitation. Comparative analyses using the orthographic word as the unit of comparison were misleading, because the word is constituted very differently in languages with different inflection systems (e.g. agglutinative languages) and word-formation processes (e.g. languages with a greater compounding tendency) (see Table 2).

Table 2. Comparing lengths of English source texts and their translations

Text	**Text Length (words/tokens)**						
	En	**Fr**	**Ge**	**Sp**			
The Laws of Football (FIFA)	10,216	11,763	9,173	11,030			
	En	**Po 1**	**Po 2**	**Sp**			
Alice in Wonderland (Carroll)*	31,731	25,348	26,245	25,566			
	En	**Bu**	**Cz**	**Es**	**Hu**	**Ro**	**Sl**
1984 (Orwell)**	104,302	87,235	80,366	79,334	81,147	101,460	91,619

Key: **En** = English, **Fr** = French, **Ge** = German, **Sp** = Spanish, **Po** = Portuguese, **Bu** = Bulgarian, **Cz** = Czech, **Es** = Estonian, **Hu** = Hungarian, **Ro** = Romanian, **Sl** = Slovene

* from de Borba (1997, 1999)

** from MULTEXT-East (Tomaz Erjavec, p.c.)

Sinclair's objections to morphemic analysis were that they were an "abstract category" that were computationally undiscoverable, and needed subjective manual annotation of the corpus prior to analysis. However, there have been some

attempts at (semi-)automatic discovery of English morphemes (e.g. Baroni 2003, Richens 2011), so corpus linguistics may yet operationalise this mode of analysis in the future.

6. Word mode

The word is probably the linguistic unit most widely recognised in all languages and by all cultures, by linguists and non-linguists. Sinclair (2004: 24) confirms its primacy: "The starting point of the description of meaning in language is the word". Naming 'word' and 'sentence' as the two 'primitives' in language form, he continues:

> the word is the unit that aligns grammar and vocabulary. The alignment of grammar and vocabulary is very clear in inflected languages, where in the typical case one morpheme, the lexical one, is invariable and the other, the inflection, varies with the local grammar.

However, as Teubert (2010: 5) says: "in spite of all attempts to pin down the accurate meaning(s) of a word, word meanings have a tendency to remain fuzzy", and "[i]ndeed, from a semantic perspective, the word as the basic unit of language has been shown to be a rather poor choice. Single words are notoriously ambiguous" (Teubert 2010: 6). Hence, as with each unit we encounter, corpus linguistics needs to proceed very cautiously.

Tokenisation (segmenting texts into words) is usually the first step in corpus analysis, but this is not a straightforward task for all languages. Written English can be tokenised using a relatively simple, computer-tractable definition of 'word', as "a sequence of characters bounded by spaces", but this obviously does not apply to spoken texts; and even with written texts, difficult decisions need to be made on how to treat punctuation marks, hyphens, and other non-alphanumeric characters. After tokenisation, orthographic conventions like capitalisation also need to be addressed before words (tokens) can be assembled into a word (types) frequency list.

Word (type) length is an aspect of English that is less studied. In the 16.78m-word written component of the BCET, the number of tokens for each type length was: 3-character types (4m tokens), followed by 4-, 2-, 5-, 6-, 7-, 8-, 1-, 9-, 10- (0.4m tokens), 11-, 12-, etc. As *the* is the most frequent word (type), it is not surprising that 3-character types constitute nearly 25% of all the tokens, or that shorter types (mostly grammar items) occur more frequently than longer (lexical/content/vocabulary) types. However, once the list was re-ordered by the number of types per type-length, it looked far less linguistically intuitive: 8-character types (28,000)

were the most frequent, followed by 7-, 9-, 6-, 10-, 11-, 5-, 12-, 4-, 13-, 14-, 3-, 15-, 16-, 17-, 18-, 2- (752), 19-, 20-, etc. (with 1-character types between 27- and 28-). The most frequent longer types were often hyphenated (*nineteenth-century* (168 tokens), *characteristically* (89), *industrialization* (88)). The longest types (up to 65 characters) occurred once each, and were mostly lengthy technical expressions, or informal utterances, often used as modifiers (*he-who-came-from-big-land-across-big-sea, co-existence-from-positions-of-strength, keynes-plus-modified-capitalism-plus-welfare-state, ask-no-questions-and-you'll-be-told-no-lies*).

Corpus linguistics assumes that the frequency of a linguistic unit or feature reflects its significance (even if we cannot be more precise on the nature of that significance) in the texts. The word type frequency list is therefore a vital first step in any analysis, whatever its purpose (language description, pedagogy, translation, stylistics, discourse analysis, computational linguistics, etc.); it often guides decisions on which items to analyse, how much effort they deserve (within the purpose), the level of detail obtainable, and so on. On the other hand, this does not mean that low-frequency features can be ignored, e.g. Larrivée and Krishnamurthy (2009) looked at the 'determiner + indefinite pronoun' feature proscribed by French and English grammars, and discovered that it occurred with sufficient frequency to allow the study of both conventional and creative usages.

The most frequent types in any large general corpus of English, whatever its size are *the, of, and, to, a, in, that, is it*, etc. (with minor variations in rank). The regular relationship between frequency and rank is usually explained with reference to Zipf's Law, which predicts that a few high frequency types will account for most of the tokens, and that there will be many low frequency types (Piantadosi 2014: 1112). Zipf's law also predicts that the 2nd ranked item will be half the frequency of the 1st, and the 3rd item will be one-third of its frequency, etc.[4] Zipf suggested that his law was applicable in many fields of human behaviour because of "the principle of least effort", and indeed Adamic and Huberman (2002: 143) apply it to city sizes, the distribution of income, and internet websites, pages, and visitors). Piantadosi (2014: 1112–3) echoes this principle in terms of speakers needing to communicate similar meanings in "a given world or social context". Montemurro (2001: 569–570) describes the Zipfian curve on a graph as "step-like plateaux", adding that the last plateau is the longest, corresponding to "hapaxes… words that appear just once in the text", and this is indeed a consistent feature in

4. However, Wyllys (1981: 55–6) adds the caveat that "the approximation is much better for the middle ranks than for the very lowest and the very highest ranks". This is also borne out by comparing the 18m-word BCET and the 418m-word *Bank of English*: *been* is ranked 48 and 47 respectively, *people* is 75 and 72, *how* is 94 and 104, and *widely* is 2500 and 2486.

corpora: whatever their size and exact composition, the proportion of hapaxes remains at c. 50% of the total types (see Table 3).

Table 3. *Bank of English* corpus: hapaxes

	Bank of English corpus				
Year	**1987**	**1993**	**1995**	**1996**	**2000**
Total tokens	18,000,000	120,000,000	211,505,963	323,302,789	418,449,873
Total types	247,069	475,633	638,901	812,452	938,914
Hapaxes	131,299	213,684	296,436	383,356	438,647
% of hapaxes	53.14%	44.93%	46.40%	47.19%	46.72%

7. Lemma mode

For highly inflected languages, it may seem efficient and practical to reduce the size of the word frequency list by considering all the inflected forms as the same word type, i.e. as members of a lemma (a root/citation form), for some analytical purposes. However, Sinclair (Sinclair 2004: 17, Sinclair, Jones & Daley 1970[2004]: xix) raises several problems. He notes that different forms of the same lemma may vary considerably in both frequency and range of meanings. Automatic lemmatisation brings additional problems: it is based on intuition-based, non-empirical dictionaries or paradigms. Nevertheless, as it is often difficult to decide beforehand which set of forms will be relevant for a specific analysis, lemmatisation can be a useful optional function: it can be informative and yield insights: e.g. unlike the word frequency list (*the, of, and, to, a, in*, etc.), the lemmatised list reveals *be* to be the 2nd most frequent lemma in English after *the*, consisting of 17 types in the *Bank of English* (see Table 4). Similarly, *have* is the 8th ranked lemma and includes 15 types.

Table 4. Lemmatised word frequency list (*Bank of English*)

Lemma	**Tag**	**Frequency breakdown**				
THE	DT	17,845,179				
BE	V	11,989,968				
	VB	*be*	1,702,992	VBR	*aren't*	20,257
	VBD	*were*	828,676	VBR	*art*	80
	VBD	*weren't*	14,366	VBR	*'re*	246,235
	VBDZ	*was*	2,423,790	VBZ	*is*	2,970,439
	VBG	*being*	274,115	VBZ	*isn't*	50,691

Table 4. (*continued*)

Lemma	Tag	Frequency breakdown				
	VBM	*am*	68,169	VBZ	*'s*	910,625
	VBM	*'m*	207,711	VBZ	*was*	13
	VBN	*been*	801,065	VBZ	*wasn't*	55,201
	VBR	*are*	1,407,720			
OF	IN		8,321,789			
AND	CC		7,582,146			
A	DT		7,060,847			
IN	IN		5,826,066			
TO	TO		5,132,283			

8. Concordance mode

Corpus linguistics asserts that the meanings, patterns of behaviour, functions, pragmatic effects, etc., of any linguistic unit can only be ascertained from its contexts of usage. The 'concordance', which shows examples of the co-text, the immediate textual context of the unit under scrutiny, is therefore a standard tool. Computers have made concordances easy to produce, but Cruden manually compiled a concordance to the Bible in the 18th century, to better understand the meanings of the biblical words from their collocations (Cruden 1769: vii; see Section 9). Concordances were later manually compiled for the works of Shakespeare, Milton, Tennyson, Homer, and Dante. Cruden used only very brief contexts, but corpus software usually offers a range of contexts, the default depending on the size of the display window, only copyright restrictions on the corpus texts limit the maximum contexts that are distributable. Concordance outputs are usually in KWIC (**K**ey **W**ord **I**n **C**ontext) format, with the node (word under scrutiny) in the centre, and a few words of context on either side.

Krishnamurthy (2005a) showed how even the first screenful of concordance examples of *surge/-s/-ing/-ed* (i.e. just 23 examples) illustrated the main usages, and enabled lexicographers to draft word senses. Sinclair (1991: 84), in his study of *of*, analysed one batch of 30 randomly selected examples, then another, and reported that very little new information was found after that. Stubbs (1995: 20) suggested that only our growing experience in corpus analysis will gradually tell us the optimal number of examples to examine for particular items.

Krishnamurthy (2001a) showed the wide range of information obtainable from concordance examples. He extracted items beginning with *receiv-* from

an alphabetical frequency list, noting the most frequent forms: *received* (48,309 tokens), *receive*, *receiving*, *receives*, *receiver*, *receivers*, *receivership*, *receivables*, *receivable*, *receiverships* (28); *receivability, receival/-s*, and *receiveth* were rare (< 5). It is unwise to make any grammatical or semantic assumptions at this stage, e.g. *receivable* may be an adjective, and/or *receivable* and *receivables* may be a countable noun. Only the concordance reveals the usage facts: both forms occur in business-related contexts; *receivable* is an adjective (unusually, used only as a postmodifier); never a singular noun; *receivables* is a plural noun:

however, have trimmed its interest	**receivable**	figure by £22
couldn't tell whether accounts	**receivable**	had been paid
offset by an increase in accounts	**receivable**.	Later, when
would routinely tell her they had	**receivables**	of thousands
dollars of credit-card	**receivables**	it has so fa
The value of debtors (or	**receivables**)	has to be p
foreign-currency payables against	**receivables**.	This type

Collocations may also suggest lexical relations, e.g. *payables* as an antonym. Intuition may offer *put down* as an antonym of *pick up*, but it is not always so with a telephone *receiver*:

a glance at Andrei. Replacing the	**receiver**,	she spoke hurr
of the dining room, picked up the	**receiver**.	'Hello," she s
want it done now!" He replaced the	**receiver**	then walked Kol

This is confirmed by the *Glowbe* corpus, which has 27 examples for *replace* and only 10 for *put down*. A semantic nuance has emerged from the concordance scrutiny: *replace* indicates a terminated call; *put down* involves a temporary interruption.

Other groups of examples display the collocations that distinguish between senses, and more lexical relations: e.g. *flights, aircraft, on-board* specify air-traffic communications; *receivers* and *transmittters* are antonyms:

the path of incoming flights. A	**receiver**	on an aircraft
room's conversations to a nearby	**receiver**	– even an FM
and causing 'static" in on-board	**receivers**	and computers.
waveband with other transmitters,	**receivers**	and communica

Whereas, *creditors, pensioners, banks, debts* indicate a business sense, and there are three collocating verbs: *appoint, bring in, call in*:

we had anticipated to appoint a	**receiver**	who may well
to creditors. The Official	**Receiver**	– part of the
Pensioners are keen to bring in the	**receivers**	to protect wha
A syndicate of banks called in the	**receivers**	over debts tot

Thus, just a brief glance at a few concordance examples can reveal word-classes, word senses, lexical relations, collocations, lexical domains, and phraseologies. This procedure can also be usefully reversed for pedagogic purposes. Deleting the node or key word in a concordance ("gapped" concordance) can induce language students to realise how much information they can obtain about an unfamiliar word or usage from the familiar elements in its co-text:

to make a sturdy coffee	-----,	with plenty of room for
others who sit round her	-----	include the business men
MACPHERSON: Now on the	-----	here we've got what looks
a loveseat, a coffee	-----,	two large cardboard
suitcases lay open on a	-----.	The silver had already
1818? [p] Answer: From	-----	C, we find the year 1816
I lowered my tray	-----	from the back of the seat
end of the dark shiny	-----	and toasted each other in
the size of a snooker	-----.	Another room upstairs is

Even many post-beginners will be able to identify the missing word as *table*. Krishnamurthy (2001b) suggested other ways to stimulate students' curiosity, using concordance examples for *accident*, asking questions, and making further corpus searches to find the answers.

9. Collocation mode

Collocation (the tendency that speakers/writers have to use certain words in close proximity to each other, in ways that cannot be predicted purely by semantics or grammar) is probably the single most interesting feature of language established and explored by corpus linguistics (see for example Krishnamurthy 1996a, 1996b, 2000, 2005b, forthcoming; Sinclair, Jones & Daley 1970 [2004], Barnbrook, Mason & Krishnamurthy 2013).

Ancient linguistics (see Section 1) said very little about the co-selection or phraseology of lexical items. Perhaps the profusion and complexity of word forms led to the focus on morphology, phonology, and (later on) syntax. Root forms

were assembled thematically or semantically (grouping near-synonyms). But, as Basham (1967: 391) noted, living languages tend to develop "in the direction of simplicity", and among the Indo-European family perhaps English has become "simpler" than the other members, retaining only a very limited inflectional system, making the formerly optional prepositional/lexical indicators obligatory, and expressing clause functions through word order rather than morphology. Firth (1957a: 196) emphasised that "[m]eaning by collocation is an abstraction at the syntagmatic level", so perhaps our attention had to shift from paradigm to syntagm before collocation could be even noticed, let alone attain significance. Sinclair, Jones and Daley (1970[20040]: 19) say: "to describe fully the collocational behaviour of even the 3,000 most common words in the language a text of several million words would be required", which suggests a more pragmatic reason for the lack of collocation studies in ancient times was simply the lack of technology.

Cruden (1769: 121) recognised the role of proximate items in creating meanings: "by the words annexed to DRY, the meaning is obvious", but the word *collocation* only acquired its current sense much later: the OED has a literary citation from 1862; Palmer (1933) used it for wordclass-combinations, e.g. 'verb + noun'; Firth (1957b: 196) first asserted its primarily lexical nature. Echoing Halliday's definition of lexis as "most delicate grammar" (Halliday 1961: 267), we could perhaps describe collocation as "most delicate context". Moon (1987: 92) offers a succinct Firth-inspired description of collocation as the "lexical realisation of the situational context". Sinclair (1991: 8) polarises grammar as "part of the management of text", and collocation as the part of the process of meaning-creation. He says that "words… do not normally constitute independent selections… the item and environment are ultimately not separable" (Sinclair 2004: 19) and that "many of the word-by-word choices in language are connected mainly syntagmatically; the paradigmatic element of their meaning is reduced to the superficial" (Sinclair 2004: 22). Collocation "illustrates the idiom principle; words are often chosen in pairs or groups, not necessarily adjacent" (Sinclair 1991: 115) and "a language user has available… a large number of semi-preconstructed phrases that constitute single choices… although [they] appear analysable into segments" (Sinclair 1991: 110). The idiom principle is "relegated to an inferior position in most current linguistics, because it does not fit the open-choice model" (Sinclair 1991: 110). Sinclair (2004: 29) sees the open choice principle as being realised in the terminological tendency (individual words acquiring fixed meanings in relation to the world) and the idiom principle in the phraseological tendency (words making meanings by their combinations).

In 1985, the only COBUILD resource for identifying collocations was a right-sorted concordance printout. Krishnamurthy (1987, 2000) reports the later provision of left-sorted concordances and extended contexts. An analysis of *crisp*

in 1992 revealed the discrepancy between dictionary and corpus collocates: (a) OALD4 (1989) listed (within 6 numbered senses): *biscuit, pastry, toast, snow, apple, lettuce, note*, *morning, air, curls, hair, order, speech, answer*; (b) the BCET collocates (using the OALD4 sense divisions, but in corpus frequency order) were: *air, morning* (20); *pastry, snow* (6); *hair, curls* (5); *lettuce* (2); *note* (2); (c) there were no examples at all of *biscuit, toast, apple, order, speech, answer*. Intuition is clearly not a reliable basis for retrieving collocation. The adjective collocates of *toast* in the corpus were *burnt, dry, burned, cold, soggy*, suggesting that we describe toast when it is unsatisfactory (i.e. not *crisp*). Sinclair (1987b: 146) suggests that "[w]hen people are asked to think of collocates, they frequently think more of semantic sets".

Collocation is also very useful in disambiguating near-synonyms: the most frequent BCET collocates for *electric* (735 occurrences) were: *light* (37), *charge*, *power*, *fire*, *lights*, *shock*, *fence*, *bulb*, *heating*, *motor*, *shocks*, *blue*, *razor* (6); and for *electrical* (262 occurrences): *system* (12), *energy*, *appliances, current*, *equipment*, *output*, *goods*, *stunning*, *apparatus*, *charge*, *devices*, *failure*, *fire*, *machinery*. The collocates of *electric* refer to individual devices that are powered by electricity, whereas the collocates of *electrical* are abstract or collective terms. Examples (from *Glowbe*) of the only shared collocate, *fire*, confirm this: … *a humming, single-bar electric fire; I turned the electric fire on*; and *razed to the ground by an electrical fire; an electrical fire triggered by a faulty exhaust* (see also Krishnamurthy 2008).

Firth (1957b: 196) uses the example of *dark night* to assert that "[o]ne of the meanings of *night* is its collocability with *dark*." Sinclair (Sinclair, Jones & Daley 1970[2004]: xxi) suggests that, in this case, "the adjective *dark* is not selecting among all possible nights, the dark ones, but is reinforcing the dark element already in *night*" and calls this the "focusing" function of an adjective (as opposed to the familiar "selective" one).

9.1 Adjacent collocations: N-grams

Collocations may consist of adjacent or non-adjacent items, but the easiest to spot are obviously the adjacent ones. Most concordance software enables this by alphabetically sorting words to the right or left of a nodeword. However, this will only indicate the adjacent collocates of a specific word. Another problem is that although two or more words may frequently occur next to each other, they may not form a new unit of meaning; frequency alone is not enough, e.g. *of the* and *in the* may be extremely frequent sequences (see below), but they are not units of meaning. Nevertheless, sequences of adjacent words in a corpus can be easily extracted by software (cf. sequences of adjacent characters in Section 4), and are usually called 'n-grams'.

The most frequent 2-grams (bigrams) are combinations of function words, e.g. in a *Times* newspaper corpus (1993): *of the* (40,683), *in the, to the, for the, on the, to be, at the* (9529). If we eliminate function words, units of meaning do emerge, e.g. in the 121m-word *Bank of English* (1993): *per cent*, *united states, prime minister, soviet union, wall street*. As well as compound nouns, we find verbal phrases (*ve got*), adverbial phrases (*too much*), and calendar dates (*september 1992*). The information obtained from n-grams can vary according to corpus composition and size, e.g. the top 1-grams in the single text *Alice in Wonderland* indicate the main character (*Alice*), the gender of that character (*she*), and the generic nature of the text: many items indicate conversation or reported speech (*said, I/you*); 2-grams confirm the dialogic nature of the text, and its main speakers (*said the, said Alice)*; 3-grams identify more characters and speakers (*the Mock Turtle, said the King, the March Hare, said the Hatter)*.

Although 4-grams and larger units may add a few more text-specific chunks, they also tend to add many more phraseologies from general English (*a minute or two*).

Mahlberg (2013) used n-grams[5] in her study of Dickens. As his work often appeared as weekly episodes, descriptive phrases (e.g. *his hands in his pockets*) were deliberately repeated in order to remind readers of the character's identity. Krishnamurthy (2013a) compared 4-grams in the election manifestos of UK political parties since 1900, to show which issues each party has focussed on. The Google Ngram Viewer[6] displays n-grams in the Google Books collection from 1500 to 2008. However, n-grams only cover adjacent collocations. Non-adjacent collocations may also be positionally fixed in some cases, e.g. *make* + [*my/your/her/their*] + *way*; *on the* + [*verge/brink/edge*] + *of collapse*.

9.2 Non-adjacent collocations: Span and statistics

When dealing with non-adjacent collocations, we need to decide how far apart the collocating items can be. Halliday (1966: 152) describes this as "some measure of significant proximity, either a scale or at least a cut-off point". Sinclair (Sinclair, Jones & Daley 1970[2004]) asserts that all words have collocational patterns. The greater frequency of grammatical words in texts entails their more frequent presence in collocations; however, their position is usually fixed. Lexical words tend to have greater positional flexibility. After experimenting with greater distances, he says that the optimal span for English is ± 4 words; "[a] shorter span would miss

5. Also called 'clusters' (Mahlberg 2013), 'lexical bundles' (Biber 2009), or chunks.

6. <https://books.google.com/ngrams>

valuable evidence", but longer spans obscure the collocation patterns with excessive irrelevant information (Sinclair, Jones & Daley 1970[2004]: 5; Sinclair 1991: 170).

It is possible that the cut-off point for collocation span actually depends on the limitations of human memory. Miller's (1956: 96) psychological experiments in attention span and immediate memory revealed "the magical number seven": we can retain seven objects in our attention span, and seven digits in immediate memory. However, Miller then discovered that we can organise digits (or other units of information) into patterns which "recode" them and allow us to remember larger chunks, especially in language. Cowan, Morey and Chen (2007) summarise subsequent research, exemplify chunking (9 letters, *USAFBICIA*, are easily remembered when recoded as 3 acronyms, *USA, FBI, CIA*), and show that "more coherent strings of words led to larger chunks recalled" (though not more chunks) (Cowan, Morey & Chen 2007: 10); word length and pronounceability also matter, as well as memorisation strategies.

Halliday (1966: 159) alerted us to the fact that "[t]he occurrence of an item in a collocational environment can only be discussed in terms of probability", and Sinclair, Jones and Daley (1970[2004]) detail their efforts to accommodate the probabilistic nature of collocation by means of various statistical methods. COBUILD used the t-score statistic for reasons explained in Clear (1995): "T-score is a measure of the confidence with which we can assert that an association exists between two events (words)… Mutual Information (MI) is a measure of the strength of association between two events". Hence t-score is more reliable with higher-frequency co-occurrences, and MI is high when rarer items in the corpus happen to co-occur. Table 5 shows that using frequency alone, any collocate list is dominated by the corpus-frequent words; t-score reduces the impact of these (e.g. *the, and, a, of, for*), retains those that collocate significantly with the node (e.g. *it, to, is, s*), and highlights significant collocates with a lower frequency (e.g. *very, find, working, worked, so*).

Table 5. Collocates of *hard* in the *Bank of English*: Ranking based on frequency and t-score

Collocate	Frequency	Collocate	t-score	Frequency
to	58,444	*it*	146.778704	37,451
the	39,225	*to*	136.672099	58,444
it	37,451	*work*	99.794844	11,378
and	27,875	*very*	73.848184	7297
a	24,531	*is*	64.938617	18,314
is	18,314	*find*	64.513063	4915

(*continued*)

Table 5. (*continued*)

Collocate	Frequency	Collocate	t-score	Frequency
s	16,947	*working*	61.992590	4347
of	14,980	*worked*	61.399611	4039
for	11,462	*s*	55.243460	16,947
work	11,378	*so*	49.255196	5902

9.3 Collocation and phraseology

Collocation is the first step in a lexical approach to phraseology. In addition to adjacent collocations (n-grams), and separable collocations (some phrasal verbs, and phrases with fixed slots and a small set of slot-fillers, e.g. *make* + [*my/your/her/their*] + *way*), there are more truly variable phrases which can involve a wide range of collocations and greater positional flexibility.

As there are very few exact synonyms, any change of item or position may cause differences in meaning or usage, e.g. a study of *on the brink of* and *on the verge of* reveals that: (i) *verge/on the verge of* are more frequent than *brink/on the brink of*;[7] (ii) *on the verge of* covers 75.48% of *verge* usages, *on the brink of* is only 44.25% of *brink* usages; (iii) 10 of the top 20 N+1 collocates are the same for both phrases, and even the non-shared collocates are negative or neutral, except for the positive phrase *on the verge of* + *victory*; (iv) in the phrases *on the* X *of collapse/extinction/bankruptcy*, over 80% of occurrences of X are *verge* or *brink*; (v) in *on the* X *of war*, *brink* is more frequent than *verge*. Halliday (1966: 156) uses the term 'lexical set' for the items involved in collocational slots in this way, and Hanks (2013) suggests that such sets can be assembled into semantic or functional groups.[8] Krishnamurthy (2005a) analyses the complexity of phraseological variations involving *knuckle/knuckles* and the lemma *rap*, which are often very difficult to summarise in dictionaries (e.g. OALD (2000): "(give sb/get) a rap on/over/across the knuckles… rap sb on/over the knuckles | rap sb's knuckles)".

9.4 Collocation and grammar

Firth discussed the regular co-occurrence of grammatical categories (i.e. word classes), terming them 'colligations'. Colligation "deals with a mutually expectant order of categories" (cited in Palmer 1968: 186). Kenny (2014: 89) says the

7. <http://corpus.byu.edu/glowbe/>

8. For more details see <http://nlp.fi.muni.cz/projekty/cpa/> and <http://pdev.org.uk> for verb phraseologies (1 July 2017).

"relationship… between a lexical item and the grammatical classes of the items in its environment, is subsumed by Firth's category of 'colligation'", adding that "Sinclair (1996: 85) calls this 'full colligation'" (Kenny 2014: fn25).

The term 'colligation' has not been used much by other scholars, but Sinclair (2004: 35–8) uses it quite often: *true feelings* colligates with a possessive adjective, *brook* colligates with nouns and modals, *my place* colligates with a preposition and an adverb of place. Bullon and Lane (1991) use a part-of-speech-tagged corpus to distinguish the collocations of *file* in its verb usages from its noun usages. Biber (2009: 275–6) distinguishes between 'multi-word lexical collocations' (content words only) and 'multi-word formulaic sequences' (function words and content words). Running the collocation "picture" programme on a part-of-speech tagged corpus obtained the following output (Table 6) for the noun *zone* (Krishnamurthy 1992a).

Table 6. The "picture" output for the colligations of *zone*

−5	−4	−3	−2	−1	ZONE	+1	+2	+3	+4	+5
P	NP	NP	D	AJ	N	VX	VX	VN	P	D
VC	N	P	D	AJ	N	WH-	VC	D	NPOS	D
TO	VF	D	NUM	N	N	#	WH-	VC	ADV	D

Key: *P* = pronoun, *NP* = proper noun, *D* = determiner, *AJ* = adjective, *VX* = auxiliary verb, *VN* = the *–ing* form of a verb, *WH-* = *wh-* word, *NPOS* = possessive noun, *VF* = finite verb, *NUM* = number, *N* = noun, # = punctuation, *ADV* = adverb

The link between colligation and meaning was systematised in the Particles Index of the *Collins COBUILD Dictionary of Phrasal Verbs* (1989), cited by Sinclair (2004: 20) in support of his assertion: "The meaning of words chosen together is independent from their independent meanings". After this, COBUILD switched from using functional labels (e.g. S for subject, O for object, etc.) to formal word-class labels (e.g. 'V n *to* n') in the 2nd edition of its dictionary (1995), and this process culminated in 'pattern grammar' (Francis, Hunston & Manning 1996, 1998, Hunston & Francis 2002), in which all the colligational patterns involving verbs, nouns, and adjectives were grouped by the meanings they generated.

9.5 Collocation and evaluation: Semantic prosody

The most exciting discovery to emerge from the focus of corpus linguistics on collocation is probably 'semantic prosody'. The idea originated in Sinclair's (1987b: 155–6) observations that the subjects of the phrasal verb *set in* generally referred to "unpleasant states of affairs" (*rot*, *decay, malaise*, etc.) and that "seemingly neutral words can be perceived with positive or negative associations through

frequent occurrences with particular collocations" (Sinclair 1991: 70–75, cited in Macarro & Peñuelas 2014: 2.3.1). Louw (1993: 157) rephrased this as "a consistent aura of meaning with which a form is imbued by its collocates" and cited further examples (*bent on, symptomatic of, utterly*, and *without feeling*). Krishnamurthy (2001c) found the most frequent collocate of *arrant* to be *nonsense*, and its other collocates to be mostly negative (*bigot, bullshit, chauvinism*, etc.), but noted some "neutral" or even "positive" words (*beginner, democracy, romanticism, Scottishness*) that became negative in its proximity. Positive prosodies were also discovered: *flexible* (Tognini-Bonelli 2001: 18–24), *important consequences* (Xiao & McEnery 2006: 113), and *provide* (Zethsen 2006: 279). Sinclair (2004: 35–8) attempted more precise labels: "reluctance" or "inability" for *true feelings*; "informal invitation" for *to my place*.

Sinclair (1996, 2004: 30–5) offers an exhaustive phraseological analysis of *naked eye* involving collocation, colligation, semantic preference, and semantic prosody. Although *naked eye* is semantically very opaque (i.e. eyes do not wear clothes), "we can… make a metaphorical extension to *naked* which fits the meaning" (i.e. "without the aid of high-powered optical instruments"), but Sinclair warns that "it is dangerously easy to reverse the procedure and assume that the metaphorical extension is obvious. It is not" and suggests other feasible meanings (e.g. unprotected, without eyelids, without spectacles). Looking at the far left context, Sinclair observes a vague feature (adjectives: *small*, *faint*, *weak*, *difficult*; adverbs: *barely*, *rarely*, *just*; modal verbs: *can, could*), which he gradually identifies in over 85% of examples as a *semantic prosody* which he labels "difficulty". Krishnamurthy (1996b) gives a similar analysis of *cutting edge*; it is mainly used in complement position after a copula verb, and has a clearly identifiable positive semantic prosody (collocates: *progressive*, *excitement*, *original*, *modern*, *exciting*).

Table 7. Collocations and colligations of *cutting edge* in the 323m-word *Bank of English*

BE	*at*	*the*	***cutting edge***	*of*	*the*
12% (=285)	17% (=425)	42% (=1022)	**(=2424)**	26% (=641)	8% (=191)
	on				*technology*
	7% (=168)				5% (=114)

Malinowski said language was a "mode of action" (see Section 2), and Firth said linguistics was "the study of 'processes and patterns of life'" (cited by Lukin 2014), using 'process' to describe "both the linguistic and non-linguistic aspects of the context of situation". Krishnamurthy (forthcoming) sees collocation as "the mechanism, process, and force that enacts language change". Meaning arises from context, and context is continually changing, so collocational changes are actually the process of meaning formation, and language change merely the sum of the

collocational changes. Krishnamurthy (2003) notes the changes in collocation from *sleazy* (a 17th century loanword used to describe places and activities in a mildly pejorative way (*unattractive, cheap, disreputable*) to *sleaze*, a c. 1980 back-formation, which soon becomes highly political (*government, party, MP*) and highly pejorative (*scandal, corruption*). More recent work on diachronic collocation, e.g. Alba-Salas (2007) and Kehoe and Gee (2009), may strengthen this view.

Perhaps collocation affects semantics, and semantic prosody affects pragmatics, especially evaluation. The more frequently that words are used, the weaker or less specific their meaning seems to become. Krishnamurthy (2002a) shows how *sexy* (first attested in c. 1890 with the negative meaning "engrossed in sex"; used positively in the 1920s, meaning "sexually attractive") has increased rapidly in usage, has mostly feminine referents (*she, her, woman, women, girl, girls, female)*, and focusses on appearance (*look, looking, looks)* of body or clothing. Even the gender-neutral/abstract nouns (*image, look, mood, voice, curves, pout*) relate to women in the concordance examples. It now also refers to art and advertising, (*film, photos, ads, campaign, play, thriller*) and products that were sexualised by 20th century advertisements (*cars, motorbikes, cigarettes)*. But the most striking feature of the data in general is actually the absence/loss of the connection with women, or indeed with sexuality at all. A large proportion of the occurrences have a general positive prosody (good, exciting, fashionable), and journalists use it to describe anything and everything (e.g. cities, foods, sports, computers, politics, science, the arts, success, careers, pensions, price stability). Frequently used words also seem to tend towards one prosodic pole (whether positive or negative), and some may even flip polarity from time to time, e.g. *funky, bad, wicked, sick*.

10. Text and corpus mode

Proceeding directly from collocation to 'text and corpus' level may seem surprising, especially in view of Sinclair's assertion that sentence is the second primitive in the description of meaning in language (see Section 6) and "aligns grammar and discourse" (Sinclair 2004: 24). However: (i) sentences are extremely difficult to identify in spoken data; (ii) many grammarians prefer clause to sentence as the main unit of analysis; (iii) McEnery and Hardie (2011: 134) argue that Sinclair's focus on sentence is a result of his use of the term 'discourse' with a meaning closer to 'text'; (iv) in most forms of writing, punctuation is increasingly inconsistent and idiosyncratic, so sentences indicate authorial intentions and style rather than linguistic units; (v) corpus software does not routinely tokenise punctuation, so sentences are less frequently analysed (unless the corpus has been automatically parsed, or manually annotated); (vi) collocation traverses sentence boundaries,

e.g. (from *Glowbe*): *try **harder**. Because it **worked** last time; It's a **hard** job, you **work** crap hours; it's **hard**, it takes **work**; I was greeted by a clear sky on Saturday **morning**. A notably **crisp** (but dry) chill originated from the south-west; Now there was a **light** inside. Not **electric***; (vii) as our ultimate focus is on meaning, it is arguable whether the sentence is a suitable unit of meaning for corpus analysis.

'Text' is as difficult to define as any other linguistic unit. Historically, linguists focussed on the sentence; text was the domain of literary and stylistics scholars, with a focus on published writing (poems, novels, plays) and authors. Some definitions of text are based on linguistic units, e.g. Werlich (1976: 23) says text is "an extended structure of syntactic units" marked by coherence and completion; Fowler (1991: 59) says text is "made up of sentences" with additional principles and rules. Carstens (2001: 589), however, invokes communicative features to identify a text: "cohesion, coherence, intentionality, acceptability, informativity, contextuality and intertextuality". Halliday and Hasan (1976: 1–2) link text directly to meaning: "any passage – spoken or written, of whatever length… a unit not of form but of meaning", a unified whole of language in use. Carter and McCarthy (2006: 926) broadly agree, but add that a text needs to be "pragmatically coherent in its real-world context". Barton and Lee (2013: 16) redefine texts for the internet age, as no longer "fixed and stable", more "multimodal and interactive", and highlight their increasing intertextuality.

Sinclair (1966) frequently refers to 'text', and even uses "a mammoth text" to mean a corpus (Sinclair, Jones & Daley 1970[2004]: 8). His final works retain this focus, e.g. "I have placed text more and more centrally during my career… the actual patterns of occurrence of words in text" (Sinclair 2007: 156). He had long denied that text was a purely literary concern (Sinclair 1965: 82), asserting the "power of linking literary text to other text" and "the exceptionally detailed common ground assumed in the minutiae of linguistic analysis" (Sinclair 1965: 88). He also emphasised that only authentic text, "a spontaneously produced, continuous stretch of natural language" (Sinclair, Jones & Daley 1970[2004]: 28) was reliable, rejecting made-up linguistic examples as evidence of people's "linguistic behaviour" rather than their "language behaviour" (Sinclair, Jones & Daley 1970[2004]: 8).

While the quantitative basis of corpus linguistics (especially its statistical tools) makes it very powerful in the analysis of large datasets, and enables us to make largescale generalisations, it is less useful for studying short individual texts (however defined). Not only are linguistic features likely to be unique or infrequent, but no generalisation is possible. Even when the complete oeuvre of a prolific author such as Dickens[9] is subjected to corpus analysis (e.g. Mahlberg

9. 4.6m-words according to <https://atkinsbookshelf.wordpress.com/2012/02/12/words-invented-by-dickens/> (1 July 2017).

2013), its features can be listed, but evaluating the features really gains robustness only in comparisons, e.g. his earlier works with his later works, his works with those of another author, or genre, or historical period.

Krishnamurthy (1995) compared Philip Larkin's poem *Spring* (1954) with the 16.78m-word written BCET and the 121m-word *Bank of English*. Despite the disparities in this comparison (the corpora contained no poetry and dated from after 1980; the poem was only 99 words long), some interesting findings did emerge (e.g. *s* was the 7th most frequent letter in the corpora but 2nd in the poem, imbuing it with sibilance; repeated character sequences produced alliteration; the first word in the poem, *green-shadowed*, did not occur in the corpora, where *green* collocated with the positive word *shade*, and never co-occurred with the negative word *shadow)*. Louw (1993) similarly analysed Larkin's poem *Days*. Krishnamurthy (1996a) compared three near-synonyms (*ethnic, racial, tribal*) in four short newspaper articles with dictionary entries and the 121m-word *Bank of English*, revealing that the words had different referents in different areas of the world, but only *tribal* was pejorative. Such techniques are now widely used in forensic linguistics (for SMS messages, suicide notes, etc.), and by political analysts (for speeches, manifestos, etc.), and in a new sub-discipline, corpus stylistics, which is rapidly emerging (e.g. Hoey 2007, Ho 2011, Mahlberg 2013).

A corpus is a collection of texts, so the corpus level of analysis logically belongs here, but as all the analyses in this paper are corpus-level analyses, it is perhaps more useful to extend the discussion of text to the nature of the corpus collection. The design of the corpus determines not only the texts included in it, but also the research questions that can be asked, the scope for extrapolation from the analyses, and the generalisability of the findings. Before computers, a corpus usually signified the complete works of an author, the extant texts from a dead language or historical period, or a collection of texts on the same topic, e.g. *the Darwinian corpus … a corpus of writing on evolution*.[10] Therefore, once computers were introduced into the process in the 1960s, linguists revised the definition of a corpus to "a collection of texts stored and analysed in a computer", and also specified their purpose as "for language description".

The *Brown Corpus* (Francis & Kucera 1979) contains 1,014,312 words from texts printed in the USA in 1961, written by native-speakers of American English. 2000-word samples were extracted from 500 prose texts (not poetry or drama), chosen as being representative of the language, not for any subjectively evaluated excellence. Although the categories, subcategories, and their sizes were decided by an academic committee, the actual texts were selected from library catalogues by random procedures.

10. <http://www.oxforddictionaries.com/definition/english/corpus>

The OSTI project (1963–70) was the first to compile a spoken corpus. Sinclair, Jones and Daley (1970[2004]: 11) describes the recording and transcription processes, and says that the small amount of data (135,000 words) was due to restrictions of time and computer technology (Sinclair, Jones and Daley 1970[2004]: 18–20). Each one-hour conversation between 3–4 people at Edinburgh and London universities yielded c. 8–10,000 words, and participants were told they would be recorded, but the topics (university life, sport, law, religion, holidays, etc.) arose spontaneously.

The *Survey of English Usage* at London University[11] started collecting samples of speech in 1959. The samples were transcribed, typed, and grammatically annotated, on paper cards. The 1m-word computerised *Survey* corpus (100 written and 100 spoken British English texts of c. 5000 words each, from 1955 to 1985) was created much later, then part-of-speech tagged, and used mainly for grammatical research and Longman reference grammars.

Several specialised corpora were created at Birmingham University in the 1970–80s, e.g. Sinclair (1987b: 1) refers to 35,000 words of classroom talk; a 1m-word corpus of Applied Science; 750,000 words of Economics texts, and corpora for language learning. Sinclair (1987b) later details the design and creation of the BCET: (i) existing digital data was requested from other universities; (ii) British Council library borrowing records were consulted; (iii) drama, poetry, and texts for/by non-adults were excluded; (iv) the corpus aimed to be 70% British, 20% American, 10% Other; 75% Male, 25% Female; and 75% Written, 25% Spoken; (v) books were scanned; newspapers, magazines and ephemera were keyed; spoken data was manually transcribed. The goal was to collect a large corpus of English language (7.3m-words in 1983, 20m-words by 1986), analyse it, and publish the analyses for students and teachers of English.

After 1986, data acquisition at COBUILD increased rapidly, due to advances in computer technology, new projects (e.g. the BBC English Dictionary), new data sources, and new copyright permissions. The goal of the *Bank of English* changed accordingly, to increase both the corpus size and the range of text types, but not to worry about fine-tuning the contents, as "even a corpus of hundreds of millions of words is of a pitifully small size in comparison to the amount of English language being generated daily" (Clear et al. 1996). However, some new problems arose in defining text, e.g. (i) newspapers and magazines arrived as one digital file per issue (daily, weekly, etc.), but could each issue be considered as one text, or should it be divided into individual articles; (ii) BBC World Service transcripts arrived as one digital file per day of broadcast output, so the same item was often broadcast more than once per day (e.g. in news bulletins), but varied in length and content. In the end, to avoid textual duplication and skewing of frequencies, only the longest

11. <http://www.ucl.ac.uk/english-usage/about/index.htm> (1 July 2017).

version of each item was added to the corpus. In many cases, rapid, pragmatic decisions had to be made, as incoming data reached enormous proportions. Teubert (2001: 142–3, 2003, 2010) also discusses various aspects of corpus text selection in relation to the purpose of the corpus.

Unlike the *a priori* design of the smaller and static Brown, SEU, and BCET corpora, Krishnamurthy (2002c) suggests that the *Bank of English* exemplifies *a posteriori* design (based on a periodic review of contents, user feedback and suggestions, and newly available resources), and is more suitable for large, dynamic corpora. The 1990s discussions about representative and balanced corpora led to efforts to standardise corpus design in multinational projects like NERC (Krishnamurthy 1992a), EAGLES, TEI, etc. The *Bank of English* is now a 650m-word subset in *Collins Corpus* of 4.5+ billion words, to which new data is added every month.[12]

11. Discourse mode

Academics define 'discourse' in many different ways, some as just another "amount of language": "one or two words as in *stop* or *no smoking*… hundreds of thousands of words in length, as some novels are. A typical piece of discourse is somewhere between these two extremes" (Celce-Murcia 2002: 123–4). Others define it as a subset of a language, determined by and determining social interactions, e.g. as "the configuration of semantic resources that the member of a culture typically associates with a situation… the meaning potential that is accessible in a given social context" (Halliday 1978: 111, cited in Fowler 1996: 7, who then adds his own definition: "a system of meanings within the culture, preexisting language"). Henry and Tator (2002: 25) say that discourse conveys "broad historical meanings", is "identified by the social conditions of its use, by who is using it and under what conditions", and that it "bridges our personal and social worlds". Kress (1985: 6–7, cited in Fowler 1996: 7) calls discourse a "systematically-organised set of statements" which expresses "the meanings and values of an institution" and structures the manner in which a topic is to be talked about.

Critical Discourse Analysis (CDA) focusses on the power relations between competing groups in social interactions, explaining how discourses "enact, confirm, legitimate, reproduce, or challenge" these relations (van Dijk 2001: 353). CDA's goal is to "investigate, reveal and clarify how power and discriminatory value are inscribed in and mediated through the linguistic system" (Caldas-Coulthard & Coulthard 1996: xi). Van Dijk (1996: 91) stresses the controls and limitations

12. <http://www.collins.co.uk/page/The+Collins+Corpus> (1 July 2017).

on access to discourse, and lists some discourse types: "everyday conversations, high school textbooks, news reports in the press, parliamentary debates, scientific discourse, and corporate discourse". Krishnamurthy's (1996a) *Ethnic, racial and tribal: The language of racism?* (see Section 10) is often cited as an important early example of the use of corpus linguistics in CDA (e.g. Mautner 2012: 32, Baker & McEnery 2015: 6), because it revealed how institutionalised patterns of usages override personal usages in newspaper articles.

Teubert (2010: 3) says discourse "turns the stuff of reality out there into objects"; only then are they "at our disposition". We cannot reliably access speakers' intentions, so the meaning of the discourse is solely "what can be found in the discourse" (Teubert 2003: 146) and "what other texts of the same discourse can contribute to its explication" (Teubert 2003: 144). The discourse "contains the beliefs, attitudes and ideas that characterise and hold together a discourse community" (Teubert 2003: 144), who collaboratively "make sense of our experiences" (Teubert 2010: 1). As well as the 'general discourse' (Teubert 2001: 142), there are 'special discourses', assembled by individuals who perceive stronger intertextual links within a subset of the texts (Teubert 2010: 120). Most texts in a discourse are "transient phenomena" and quickly disappear; however, a few texts become "key texts" and influence subsequent texts, sometimes for decades or longer (Teubert 2001: 142–3).

Teubert is critical of CDA (which sees discourse as a social practice), because for him it is discourse that constructs society and social structures (Teubert 2010: 120–1). We cannot look at society from the outside, as biologists do with ants (Teubert 2010: 120–1), and discourse is "not a mechanism… delivering predictable results, telling us what can be said and what not" but "a system that keeps creating itself… like Darwinian evolution" (Teubert 2010: 11). Linguists are part of the discourse community, but are neither "privileged" members nor "experts in meaning or knowledge" (Teubert 2010: 8). The "discourse community… is in charge of the language", and establishes its conventions (Teubert 2007: 9).

Jaworska and Krishnamurthy (2012) looked at *feminism/Feminismus* in large, extant, general corpora of English and German, and found that *Feminismus* had a far lower occurrence-rate. But only the German corpus highlighted feminists' cultural contributions (academic, artistic, religious, social). Both corpora stressed the radical element in feminist politics, and feminism's historical (rather than current) significance. Specific feminist issues (inequality, pay gap, work/childcare dichotomy) were rarely mentioned. These findings were compared with sociological research, which used much smaller, English-only datasets, and a priori rather than data-generated categories.

Grundmann and Krishnamurthy (2010) created corpora of 600,000 news articles (400+m-words) from a wide range of sources in British and American

English, French, and German from 1980 to 2007, using several search terms relating to *climate change*. Previous research had been mainly sociological, focussed on English, and much narrower in scope and timeframe. This study showed similar patterns of media attention in all four corpora, increasing rapidly after 2005. Europe expressed a greater sense of urgency and often referred to USA, whereas USA was mainly self-referential. Europe used a range of terms, but USA mainly used one: *greenhouse effect* until 1989, then *global warming* (President Bush was persuaded in 2003 to switch to *global warming*; this had become the dominant term in UK in 2004, but not in USA until 2009). In French, *effet de serre* ('greenhouse effect') remained the preferred term throughout, and in German, *Klimawandel* ('climate change') and *Klimaschutz* ('climate protection'; a term newly-discovered by the corpus analysis) rose to 1st rank in 2002. *Global warming* and its translation equivalents attracted more dramatising collocates (*threat, action, fight)* than *climate change*, except in UK; and French and German generally used more dramatising words than USA and UK.

Krishnamurthy (2013b) created a corpus of UK newspaper articles surrounding events (termed *riots* by the journalists) which took place over a few days from 6th August 2011. The data showed that UK newspapers rarely covered *riots* in other countries, and that the usage of *riots* outside this brief period related to non-UK artistic/cultural (not socio-political) events (*festival, play, theatre*). In the articles covering the UK events of 2011, *police* was more significant than *rioters, looters, youths*. The articles from the period immediately after the events gradually shifted their attention from the perpetrators, locations (*London, Birmingham, night, streets)*, and event details (*looting, shops, violence, fire)*, to the subsequent legal proceedings, the evidence (*video, report)*, statements and interviews (*I, you)*, mitigations (*homeless*), and outcomes (*prison)*. After peaking on August 11th, the coverage rapidly decreased to pre-6th levels by the 31st.

12. Conclusions

This paper discussed units of language, patterns, functions, processes and meanings, looked briefly at historical linguistic ideas and approaches, then explored in more detail some of the levels of language that have been studied by corpus linguists. The solely theoretical and qualitative approaches that dominated language studies for centuries enjoyed the benefit of great freedom of opinion, as any claim could be validated by anecdotal evidence alone. Once technological advances enabled a quantitative methodology, every claim made by theoretical, experimental, or qualitative methods could be re-examined empirically. The main problem for the quantitative approach of corpus linguistics is that, for any linguistic feature

postulated in the past, an appropriate analytical level had to be established, and computable empirical units identified, so some aspects of this problem have been discussed.

Meaning itself cannot be a **mode** of analysis, as meaning arises from context, and the full context of any piece of language is too extensive to be tractable, and constantly expands as time passes, e.g. Tagg (2014: 1) refers to Dell Hymes and "the dynamic nature of context". Meaning is, therefore, a result of analysis and is affected by the analytical process, methodology, and outcomes. Meaning is always provisional, and merely a glimpse into the dynamic process of meaning-creation. McEnery and Wilson (2001) say that corpus linguistics offers an approach to semantics which is objective, and takes account of indeterminacy and gradience. Mindt (1991) argues that meanings in texts have "characteristic observable contexts", and close scrutiny can lead to "empirical objective indicators", but categories remain fuzzy, and categorial membership depends on the frequency of the exponents (see also Pustejovsky, Bergler & Anick 1993, Stubbs 2001, Baroni & Lenci 2010).

Functional or social meaning, or pragmatics, is currently beyond automation and requires subjective, manual annotation of the data. Most words, phrases, and utterances can be deemed to have a pragmatic purpose, and are often pragmatically polyfunctional, semantic prosody (discussed earlier) is one such function. Krishnamurthy (2002c) notes that pragmatic usages are easier to distinguish in corpus concordances (see also Vaughan & Clancy 2013; Aijmer & Rühlemann 2014). Metaphor is another aspect of semantics that usually requires the annotation of corpus data (see also Deignan 2005; Stefanowitsch & Gries 2008).[13]

Multilingual (comparable or parallel) corpora are used in contrastive studies and translation studies, but do not constitute a separate mode of analysis because they merely use the same analytical modes as monolingual corpora (e.g. Martin et al. (2003) on bilingual children speaking in Panjabi and English; Grundmann and Krishnamurthy (2010) on climate change discourse in English, French and German newspapers; the GeWiss project (2009–2012) on spoken academic discourse in German, English and Polish; the Comenego project (ongoing) on business corpora in French and Spanish).[14]

Teubert (2007: 12) suggests that "the next focus of corpus linguists will be on the diachronic continuity and uniqueness of meaning of a lexical item within the

13. Deignan and Potter (2004) found 92/1000 examples for *mouth* in English meaning speech, and 102/1000 examples for *bocca* in Italian with the same meaning.

14. GeWiss project, see <https://gewiss.uni-leipzig.de/index.php?id=about_gewiss&L=1> and Comenego project, see <http://dti.ua.es/en/comenego/comenego-multilingual-corpus-of-business-and-economics.html> (1July 2017).

history of the discourse". While there has been some recent progress in the development of diachronic corpora,[15] much more clearly needs to be done in this area. However, an even more exciting development may be in multimodal corpora, the most recent innovation in corpus linguistics, which may actually require new analytical modes to be devised. As Abuczki and Ghazaleh (2013: 87) say: "Spoken corpora with transcripts alone are not sufficient for uncovering the nonverbal-visual aspects of interaction since 'the reflexivity of gesture, movement and setting is difficult to express in a transcript' (Saferstein 2004: 213)". Knight (2011: 392) cites a definition of a multimodal corpus: "'an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language… based on recorded human behaviour' (Foster and Oberlander, 2007, p. 307–308)" and explains "[T]he integration of textual, audio and video records of communicative events in multimodal corpora provides a platform for the exploration of a range of lexical, prosodic and gestural features and for investigations of the ways in which these features interact in real-life discourse" (Knight 2011: 392) (see also Allwood 2008, Bateman 2012).

References

Abramova, E., Fernández, R. & Sangati, F. 2013. Automatic labeling of phonesthemic senses. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (eds), 1696–1701. Austin TX: Cognitive Science Society. <http://staff.science.uva.nl/~raquel/papers/2013/phonesthemes_cogsci2013.pdf> (1 July 2017).

Abuczki, A. & Ghazaleh, E. B. 2013. An overview of multimodal corpora, annotation tools and schemes. *Argumentum* 9: 86–98. <http://argumentum.unideb.hu/2013-anyagok/kulonszam/01_abuczkia_esfandiari_baiat.pdf> (1 July 2017))

Adamic, L. A. & Huberman, B. A. 2002. Zipf's law and the Internet. *Glottometrics* 3: 143–150.

Aijmer, K. & Rühlemann, C. (eds). 2014. *Corpus Pragmatics. A Handbook*. Cambridge: CUP.

Alba-Salas, J. 2007. On the life and death of a collocation. A corpus-based diachronic study of *dar miedo/hacer miedo*-type structures in Spanish. *Diachronica* 24(2): 207–252. https://doi.org/10.1075/dia.24.2.02alb

Allwood, J. 2008. Multimodal corpora. In *Corpus Linguistics. An International Handbook*, A. Lüdeling & M. Kytö (eds), 207–225. Berlin: Mouton de Gruyter.

Atkins, B. T. S. & Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: OUP.

Atkinson, Q. D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027): 346–349. https://doi.org/10.1126/science.1199295

15. E.g. The *Corpus of Historical American English* (COHA), at <http://corpus.byu.edu/coha/>

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 233–250. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.15bak

Baker, P. & McEnery, A. 2015. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9781137431738

Barnbrook, G., Mason, O. & Krishnamurthy, R. 2013. *Collocation: Applications and Implications*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9781137297242

Baroni, M. 2003. Distribution-driven morpheme discovery: A computational/experimental study. In *Yearbook of Morphology*, G. Booij & J. van Marle (eds), 213–228. Dordrecht: Springer.

Baroni, M. & Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4): 673–721. https://doi.org/10.1162/coli_a_00016

Barton, D. & Lee, C. 2013. *Language Online: Investigating Digital Texts and Practices*. London: Routledge.

Basham, A. L. 1967. *The Wonder That Was India*, 3rd ed. London: Sidgwick and Jackson.

Bateman, J. A. 2012. Multimodal corpus-based approaches. In *The Encyclopedia of Applied Linguistics*, C. A. Chapelle (ed.), 3983–3991. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0812

Bergen, B. K. 2004. The psychological reality of phonaesthemes. *Language* 80: 290–311. <https://muse.jhu.edu/article/169798> (1 July 2017).

Biber, D. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Bloomfield, L. 1933. *Language*. New York NY: Holt.

Bolinger D. 1968. *Aspects of Language*. New York NY: Harcourt Brace Jovanovich.

Bradbury, J. (ed.). 1991. *Cobuild English Guides, 2: Word Formation*. London: HarperCollins.

Bullon, S. & Lane, T. 1991. The main points of the news: World service data for a world service dictionary. Listed but unpublished in *Using Corpora*, Proceedings of the 7th Annual Conference. Waterloo: UW Centre for the New OED and Text Research/Oxford: OUP.

Caldas-Coulthard, C. R. & Coulthard, M. (eds). 1996. *Readings in Critical Discourse Analysis*. London: Routledge.

Carstens, W. A. M. 1999[2001]. Text linguistics: Relevant linguistics? In *Poetics, Linguistics and History: Discourses of War and Conflict. PALA Conference Papers 1999*, I. Bierman & A. L. Combrink (eds), 588–595. Potchefstroom: Potchefstroom University. <http://www.pala.ac.uk/uploads/2/5/1/0/25105678/carstens.pdf> (1 July 2017).

Carter, R. & McCarthy, M. 2006. *Cambridge Grammar of English*. Cambridge: CUP.

Celce-Murcia, M. 2002. Why it makes sense to teach grammar in context and through discourse. In *New Perspectives on Grammar Teaching in Second Language Classrooms*, E. Hinkel & S. Fotos (eds), 121–136. Mahwah NJ: Lawrence Erlbaum Associates.

Chomsky, N. 1968. *Language and Mind*. New York NY: Harcourt Brace Jovanovich. <https://www.marxists.org/reference/subject/philosophy/works/us/chomsky.htm> (1 July 2017). (One of the six lectures is reproduced here; transcribed in 1998 by A. Blunden; proofed and corrected February 2005). https://doi.org/10.1037/e400082009-004

Church, K. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology* 6(5): 1–27.

Clear, J. 1995. Corpora: T-score in collocational analysis. Posted to Corpora-List on 12 December 1999. <http://nora.hd.uib.no/corpora/2000-2/0061.html> (the link no longer valid).

Clear, J., Fox, G. L., Francis, G., Krishnamurthy, R. & Moon, R. 1996. Cobuild: The state of the art. *International Journal of Corpus Linguistics* 1(2): 303–314. https://doi.org/10.1075/ijcl.1.2.08cle

Collins Cobuild Dictionary of Phrasal Verbs. 1989. J. Sinclair (ed.). London: Collins ELT.

Cowan, N., Morey, C. C. & Chen, Z. 2007. The legend of the magical number seven. In *Tall Tales about the Mind & Brain: Separating Facts from Fiction*, S. Della Sala (ed.), 45–59. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780198568773.003.0005

Cruden, A. 1769. *A Complete Concordance to the Old and New Testament* (1891 ed.). London: Frederick Warne and Co.

De Beaugrande, R. 1991. *Linguistic Theory: The Discourse of Fundamental Work, Section 8: John Rupert Firth*. <http://www.beaugrande.com/LINGTHERFirth.htm> (1 July 2017).

De Borba, M. C. S. 1997. Two Brazilian-Portuguese translations of wordplay in *Alice's Adventures In Wonderland*. *Cadernos de Tradução* 2: 115–126. Florianópolis: Universidade Federal de Santa Catarina.

De Borba, M. C. S. 1999. Text diversity, intertextuality and parodies in Wonderland. *Fragmentos* 16: 15–22. Florianópolis: Universidade Federal de Santa Catarina.

Deignan, A. 2005. *Metaphor and Corpus Linguistics* [Converging Evidence in Language and Communication Research 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/celcr.6

Deignan, A. & Potter, L. 2004. A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics* 36: 1231–1252. https://doi.org/10.1016/j.pragma.2003.10.010

Firth, J. R. 1930. *Speech*. London: Oxford University Press.

Firth, J. R. 1957a. *Papers in Linguistics (1934–1951)*. Oxford: OUP.

Firth, J. R. 1957b[1951]. Modes of Meaning. In *Papers in Linguistics 1934–1951*, J. R. Firth, 190–215. London: OUP.

Fowler, H. N. 1921. *Volume 12 of Plato in Twelve Volumes*. Cambridge MA: Harvard University Press. <http://www.perseus.tufts.edu/hopper/text?doc=plat.+crat.+383a> (1 July 2017).

Fowler, R. 1991. *Language in the News. Discourse and Ideology in the Press*. London: Routledge and Kegan Paul.

Fowler, R. 1996. On critical linguistics. In *Texts and Practices: Readings in Critical Discourse Analysis*, C. R. Caldas-Coulthard & M. Coulthard (eds), 3–14. London: Routledge.

Foster, M. E. & Oberlander, J. 2007. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation* 41(3–4): 305–323. https://doi.org/10.1007/s10579-007-9055-3

Francis, G., Hunston, S. & Manning, E. 1996. *Collins COBUILD Grammar Patterns, 1: Verbs*. London: HarperCollins.

Francis, G., Hunston, S. & Manning, E. 1998. *Collins COBUILD Grammar Patterns, 2: Nouns and Adjectives*. London: HarperCollins.

Francis, W. N. & Kucera, H. 1979. *Brown Corpus Manual*. Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers. Providence RI: Brown University. <http://clu.uni.no/icame/manuals> (17 September 2018).

Godart-Wendling, B. 2014. L'hypothèse de Firth: Wittgenstein, héritier de Malinowski? *Historiographia Linguistica* 41(1): 79–108. <> (1 July 2017). https://doi.org/https://doi.org/10.1075/hl.41.1.03god

Grundmann, R. & Krishnamurthy, R. 2010. The discourse of climate change: A corpus-based approach. *Critical Approaches to Discourse Analysis Across Disciplines (CADAAD) Journal* 4(2): 125–146.

Halliday, M. A. K. 1961. Categories of a theory of grammar. *Word* 17(3): 241–292. https://doi.org/10.1080/00437956.1961.11659756

Halliday, M. A. K. 1966. Lexis as a linguistic level. In *In Memory of J.R. Firth*, C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (eds), 150–161. London: Longman.

Halliday, M. A. K. 1978. *Language as Social Semiotic*. London: Edward Arnold.

Halliday, M. A. K. 2006. *Language of Early Childhood*. London: A & C Black.

Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge MA: The MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Harris, R. 1987. *The Language Machine*. London: Duckworth.

Henry, F. & Tator, C. 2002. *Discourses of Domination*. Toronto: University of Toronto Press. https://doi.org/10.3138/9781442673946

Ho, Y. 2011. *Corpus Stylistics in Principles and Practice: A Stylistic Exploration of John Fowles' The Magus*. London: Bloomsbury.

Hoey, M. 2007. *Text, Discourse and Corpora: Theory and Analysis*. London: A & C Black.

Holden, C. 2004. The Origin of Speech. *Science* 303: 1316–1319. https://doi.org/10.1126/science.303.5662.1316

Hunston, S. & Francis, G. 2002. *Pattern Grammar* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins.

Jaworska, S. & Krishnamurthy, R. 2012. On the F-word: A corpus-based analysis of the media representation of feminism in English and German newspapers, 1990–2009. *Discourse & Society* 23(4): 401–431. https://doi.org/10.1177/0957926512441113

Kehoe, A. & Gee, M. 2009. Weaving web data into a diachronic corpus patchwork. In *Corpus Linguistics: Refinements and Reassessments*, A. Renouf & A. Kehoe (eds), 255–279. Amsterdam: Rodopi. https://doi.org/10.1163/9789042025981_015

Kenny, D. 2014. *Lexis and Creativity in Translation: A Corpus Based Approach*. Abingdon: Routledge.

Knight, D. 2011. The future of multimodal corpora. *Revista Brasileira de Linguistica Aplicada* 11(2): 391–415. https://doi.org/10.1590/S1984-63982011000200006

Kress, G. 1985. *Linguistic Processes in Sociocultural Practice*. Victoria: Deakin University Press.

Krishnamurthy, R. 1987. The process of compilation. In *Looking Up: An Account of the COBUILD Project in Lexical Computing*, J. M. Sinclair (ed.), 62–85. London: Collins ELT.

Krishnamurthy, R. 1992a. *Introductory Workshops on Dictionaries*. <http://www.academia.edu/7569279/1992-COBUILD_Introductory_Workshops_on_Dictionaries_at_Brighton_University> (1 July 2017).

Krishnamurthy, R. 1992b. Data collection. NERC-WP6/WP7-57, Working Paper for EC Project: Network of European Reference Corpora. Pisa: ILC.

Krishnamurthy, R. 1995. The macrocosm and the microcosm: The corpus and the text. In *Linguistic Approaches to Literature: Papers in Literary Stylistics*, J. Payne (ed.), 1–17. Birmingham: University of Birmingham.

Krishnamurthy, R. (1996a). Ethnic, racial and tribal: The language of racism? In *Texts and Practices: Readings in Critical Discourse Analysis*, C. R. Caldas-Coulthard & M. Coulthard (eds), 129–149. London: Routledge. (Reprinted in Teubert, W. & Krishnamurthy, R. (eds). 2007. *Corpus Linguistics: Critical Concepts in Linguistics*, 179–200. London: Routledge).

Krishnamurthy, R. 1996b. The data is the dictionary: Corpus at the cutting edge of lexicography. In *Papers in Computational Lexicography, COMPLEX'96*, F. Kiefer, G. Kiss & J. Pajzs (eds), 117–144. Budapest: Hungarian Academy of Sciences.

Krishnamurthy, R. 2000. Collocation: From silly ass to lexical sets. In *Words in Context: A Tribute to John Sinclair on his Retirement*, C. Heffer & H. Sauntson (eds). Birmingham: University of Birmingham.

Krishnamurthy, R. 2001a. Language corpora: How can teachers and students use these valuable new resources? In *Selected Papers from the 10th International Symposium on English Teaching*, 59–65. Taipei: ETA/ROC.

Krishnamurthy, R. 2001b. Learning and teaching through context – a data-driven approach. *TESOL Spain Newsletter* 24, 9–10. <http://www.developingteachers.com/articles_tchtraining/corpora2_ramesh.htm> (1 July 2017).

Krishnamurthy, R. 2001c. The science and technology of corpus, and corpus for science and technology. In *La Investigacion en Lenguas Aplicadas: Enfoque Multidisciplinar*, G. A. De Cea & P. D. Escribano (eds), 79–114. Madrid: Fundacion Gomez Pardo & Universidad Politecnica de Madrid.

Krishnamurthy, R. 2002a. The corpus revolution in EFL dictionaries and Appendix: Analysis of *sexy* in the 450-million-word Bank of English corpus. *Kernerman Dictionary News* 10. <http://www.kdictionaries.com/index.html#news> (1 July 2017).

Krishnamurthy, R. 2002b. The Bank of English past, present, and future: Corpus size, composition, annotation, and software. Presented at the 2nd ILASH Half-Day Workshop on Computational Language Resources, University of Sheffield. <https://www.academia.edu/7528217/2002-The_Bank_of_English_past_present_and_future_corpus_size_composition_annotation_and_software> (1 July 2017).

Krishnamurthy, R. 2002c. Pragmatics and the EFL Dictionary. Presentation at the 22nd ThaiTESOL conference, 'Inspiring Change In ELT', in Chiangmai. <https://www.academia.edu/7672953/2002-Pragmatics_and_the_EFL_Dictionary> (1 July 2017).

Krishnamurthy, R. 2003. Freeze-frame pictures: micro-diachronic variations in synchronic corpora. In *Studies in English Theoretical and Applied Linguistics*, J. Andor, J. Horvath & M. Nikolov (eds), 15–31. Pécsi Tudományegyetem: Lingua Franca Csoport.

Krishnamurthy, R. 2005a. *Grammar and Lexis of English*. MA module for Aston University. <https://www.academia.edu/7690779/2005-Grammar_and_Lexis_of_English_GLE_> (1 July 2017).

Krishnamurthy, R. 2005b. Teaching and learning English metaphors. Presentation at JALT Conference, Shizuoka. <https://www.academia.edu/7528355/2005-Teaching_and_learning_English_metaphors> (1 July 2017).

Krishnamurthy, R. 2008. ACORN in USE: CASE STUDIES. Talk given in research seminar at Aston University. <http://acorn.aston.ac.uk/RK-publications/ACORN-in-USE-CaseStudies.pdf>

Krishnamurthy, R. 2013a. Corpus linguistics: From lexis to discourse. Presentation at Hildesheim University, Germany<https://www.academia.edu/7530935/2013-Corpus_Linguistics_from_lexis_to_discourse> (1 July 2017).

Krishnamurthy, R. 2013b. 2013-Corpus Workshop III: Introduction to WordSmith Tools 6 [version 2: Using UK news articles on "UK Riots" as data]. Presentation given at 'Towards Operationalizing Corpus Development Plan', University of KwaZulu-Natal, South Africa. <https://www.academia.edu/7538315/2013-Corpus_Workshop_III_Introduction_to_WordSmith_Tools_6_version_2_using_UK_news_articles_on_UK_Riots_as_data_> (1 July 2017).

Krishnamurthy, R. Forthcoming. Collocations and lexicography: Sinclairian theory in practice. In *International Handbook of Lexis and Lexicography, Section 2: Lexical Theory and Lexicography*, P. Hanks & G-M. de Schryver (eds). Heidelberg: Springer.

Larrivée, P. & Krishnamurthy, R. 2009. La créativité et la conventionnalité de groupes nominaux atypiques déterminant + pronom indéfini et leurs contextes communicatifs. *La langue en contexte. Actes du colloque «Représentations du sens linguistique IV». Société Néophilologique de Helsinki*, 93–106.

Li, C. 2010. Unity and Variety: A Study of the Chinese Language and Its Cultural Implications. PhD dissertation, UCSD. <https://escholarship.org/uc/item/8tt1c8sd> (1 July 2017))

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 157–176. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.11lou

Louw, B. 2000. Contextual prosodic theory: Bringing semantic prosodies to life. In *Words In Context: A Tribute to John Sinclair on His Retirement*, C. Heffer & H. Sauntson (eds), Birmingham: University of Birmingham. <http://www.revue-texto.net/docannexe/file/124/louw_prosodie.pdf> (1 July 2017).

Lukin, A. 2014. The study of "living language": The SFL conception of text/context relations, 2. *Firth: J.R. Firth and renewal of connection with the processes and patterns of life*. <http://annabellelukin.edublogs.org/reading-firth/> (1 July 2017).

Macarro, A. S. & Peñuelas, A. B. C. (eds). 2014. *New Insights into Gendered Discursive Practices: Language, Gender and Identity Construction*. Valencia: Valencia University Press.

Mahlberg, M. 2013. *Corpus Stylistics and Dickens's Fiction*. London: Routledge.

Malinowski, B. 1922[1978]. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea* [Studies in Economics and Political Science 65]. London: Routledge and Kegan Paul. <https://archive.org/details/argonautsofthewe032976mbp> (1 July 2017).

Martin, D., Krishnamurthy, R., Bhardwaj, M. & Charles, R. 2003. Language change in young Panjabi/English children: Implications for bilingual language assessment. *Child Language Teaching and Therapy* 19(3): 245–265. https://doi.org/10.1191/0265659003ct2540a

Matilal, B. K. 1990. *The Word and The World. India's Contribution to the Study of Language*. Oxford: OUP.

Mautner, G. 2012. Corpora and critical discourse analysis. In *Contemporary Corpus Linguistics*, P. Baker (ed.), 32–46. London: A & C Black.

McEnery, T. & Hardie, A. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511981395

McEnery, T. & Wilson, A. 2001. *Corpus Linguistics. An Introduction*. Edinburgh: EUP.

McKeon, R. 1946. Aristotle's conception of language and the arts of language. *Classical Philology* 41(4): 193–206. https://doi.org/10.1086/362975

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81–97. https://doi.org/10.1037/h0043158

Mindt, D. 1991. Syntactic evidence for semantic distinctions in English. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, K. Aijmer & B. Altenberg (eds), 182–196. London: Longman.

Montemurro, M. A. 2001. Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A* 300: 567–578. https://doi.org/10.1016/S0378-4371(01)00355-7

Moon, R. 1987. The analysis of meaning. In *Looking Up. An Account of the COBUILD Project in Lexical Computing*, J. Sinclair (ed.), 86–103. London: Collins ELT.

Müller, F. M. 1860. *A History of Ancient Sanskrit Literature*. <http://books.google.co.uk/books?hl=en&lr=&id=cHCe48QSZaUC&oi=fnd&pg=PR1&dq=yaska+nirukta&ots=9hl1wIBfsz&sig=kf8xA2rUeWchSa-Z4q_fy5fwTI4#v=onepage&q=yaska%20nirukta&f=false> (1 July 2017).

OALD. 2000. *Oxford Advanced Learner's Dictionary of Current English*, 6th ed., A. S. Hornby & S. Wehmeier (eds). Oxford: OUP.

OALD4. 1989. *Oxford Advanced Learner's Dictionary of English*, 4th ed., A. S. Hornby & A. P. Cowie (eds). Oxford: OUP.

OED. 1971. *Oxford English Dictionary*, compact edition. Oxford: OUP.

Otis, K. & Sagi, E. 2008. Phonaesthemes: A corpora-based analysis. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, B. C. Love, K. McRae & V. M. Sloutsky (eds), 65–70. Austin TX: Cognitive Science Society.

Palmer, H. 1933. *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Palmer, F. R. (ed.). 1968. *Selected Papers of J.R. Firth*. London: Longman.

Payne, J. 1995. *Collins Cobuild English Guides, 8: Spelling*. London: HarperCollins.

Piantadosi, S. T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21: 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Pustejovsky, J., Bergler, S. & Anick, P. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics* 19(2): 331–358.

Richens, T. 2011. Lexical Database Enrichment Through Semi-Automated Morphological Analysis. PhD dissertation, Aston University. <http://eprints.aston.ac.uk/15809/> (1 July 2017).

Robins, R. H. 1964. *General Linguistics. An Introductory Survey*. Harlow: Longman, Green.

Robins, R. H. 2013. *A Short History of Linguistics*. Abingdon: Routledge.

Saferstein, B. 2004. Digital technology and methodological adaption: Text on video as a resource for analytical reflexivity. *Journal of Applied Linguistics* 1(2): 197–223. https://doi.org/10.1558/japl.2004.1.2.197

Schmandt-Besserat, D. 1977. The earliest precursor of writing. *Scientific American* 238(6): 50–58. <http://en.finaly.org/index.php/The_earliest_precursor_of_writing> (1 July 2017).

Schmandt-Besserat, D. 1991. Two precursors of writing: plain and complex tokens. In *The Origins of Writing*, W. M. Senner (ed.), 27–41. Lincoln NE: University of Nebraska Press. <http://en.finaly.org/index.php/Two_precursors_of_writing:_plain_and_complex_tokens> (1 July 2017).

Sinclair, J. M. 1965. When is a poem like a sunset? *A Review of English Literature* 6(2): 76–91.

Sinclair, J. M. 1966. Beginning the study of lexis. In *In Memory of J.R. Firth*, C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (eds), 410–430. London: Longman.

Sinclair, J. M. (ed.). 1987a. *Collins Cobuild English Language Dictionary*. London: Collins.

Sinclair, J. M. (ed.). 1987b. *Looking Up. An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. M. 1996. The search for units of meaning. *Textus* 9: 75–106.

Sinclair, J. M. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. M. 2007. Preface. *International Journal of Corpus Linguistics* 12(2): 155–157. https://doi.org/10.1075/ijcl.12.2.03sin

Sinclair J., Jones, S. & Daley, R. 1970[2004]. *English Lexical Studies: Report to OSTI on Project C/LP/08*. Re-published as Krishnamurthy, R. (ed.). 2004. *English Collocation Studies: The OSTI Report*. London: Continuum.

Stefanowitsch, A. & Gries, S. T. (eds). 2008. *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin: De Gruyter Mouton.

Stubbs, M. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1): 23–55. (Reprinted in Teubert, W. & Krishnamurthy, R. (eds). 2007. *Corpus Linguistics: Critical Concepts in Linguistics*, 166–193. London: Routledge).

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

Tagg, C. 2014. Translanguaging as an addressivity strategy for identity and relational work on Facebook. Talk given at 'Superdiversity: Theory, Method and Practice in an Era of Change' held by IRiS, University of Birmingham, 23–25 June. <https://www.academia.edu/9774990/Translanguaging_as_an_addressivity_strategy_for_identity_and_relational_work_on_Facebook> (1 July 2017).

Teubert, W. 2001. A province of a federal superstate ruled by an unelected bureaucracy: Keywords of the Euro-Sceptic discourse in Britain. In *Attitudes Towards Europe*, C. Good, A. Musolff, P. Points & R. Wittlinger (eds), 45–86. Abingdon: Ashgate. (Reprinted in Teubert, W. & Krishnamurthy, R. (eds). 2007. *Corpus Linguistics: Critical Concepts in Linguistics*, 142–178. London: Routledge).

Teubert, W. 2003. Writing, hermeneutics, and corpus linguistics. *Logos and Language* 42: 1–17. (Reprinted in Teubert, W. & Krishnamurthy, R. (eds). 2007. *Corpus Linguistics: Critical Concepts in Linguistics*, 134–159. London: Routledge).

Teubert, W. 2007. General introduction. In *Corpus Linguistics: Critical Concepts in Linguistics*, W. Teubert & R. Krishnamurthy (eds), 1–38. London: Routledge.

Teubert, W. 2010. *Meaning, Discourse and Society*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511770852

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Toury, G. 1980. *In Search of a Theory of Translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. 1995. *Descriptive Translation Studies – and Beyond* [Benjamins Translation Library 4]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4

van Dijk, T. A. 1996. Discourse, power and access. In *Texts and Practices: Readings in Critical Discourse Analysis*, C. R. Caldas-Coulthard & M. Coulthard (eds), 84–106. London: Routledge.

van Dijk, T. 2001. Critical discourse analysis. In *Handbook of Discourse Analysis*, D. Tannen, D. Schiffrin & H. Hamilton (eds), 352–371. Oxford: Blackwell.

Vaughan, E. & Clancy, B. 2013. Small corpora and pragmatics. In *Yearbook of Corpus Linguistics and Pragmatics*, Vol. 1, J. Romero-Trillo (ed.), 53–73. Dordrecht: Springer.

Wallis, J. 1699. *Grammar of English Language* (5th Ed.). Oxford: L. Lichfield.

Wang, Q. E. 2008. Beyond East and West: Antiquarianism, evidential learning, and global trends in historical study. *Journal of World History* 19(4): 489–519. https://doi.org/10.1353/jwh.0.0024

Weisser, M. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Hoboken NJ: Wiley-Blackwell. https://doi.org/10.1002/9781119180180

Werlich, E. 1976. *A Text Grammar of English*. Heidelberg: Quelle & Meyer.

Wilkins, D. 1972. *Linguistics in Language Teaching*. London: Edward Arnold.
Willis, D. 2003. *Rules, Patterns and Words. Grammar and Lexis in English Language Teaching*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511733000
Wyllys, R. E. 1981. Empirical and theoretical bases of Zipf's Law. *Library Trends* 30(1): 53–64.
Xiao, R. & McEnery, T. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1): 103–129. https://doi.org/10.1093/applin/ami045
Zethsen, K. K. 2006. Semantic prosody: Creating awareness about a versatile tool. *Tidsskrift for Sprogforskning* 4(1–2): 275–294. https://doi.org/10.7146/tfs.v4i1.324
Zhu, L. 2013. *Historical Chinese phonology as a Meeting Ground for the Indian, the Chinese, and the Western Linguistic Tradition*. <http://hiphilangsci.net/2013/09/04/historical-chinese-phonology-as-a-meeting-ground-for-the-chinese-the-indian-and-the-western-linguistic-tradition/> (1 July 2017).

Keywords

Signposts to objectivity?

Paul Baker
Lancaster University

This chapter focuses on describing, illustrating and critiquing the keywords technique, which is used to automatically identify lexical salience when comparing multiple corpora. Arguably, keywords present researchers with words that they may not have chosen to analyse in advance, thus helping to reduce researcher subjectivity.[1] I illustrate how the identification of keywords enables researchers to embark on interesting research journeys, through examples taken from an analysis of the representation of Islam and Muslims in a corpus of British newspaper articles. However, when using large corpora, even with high cut-off points for statistical salience, hundreds of keywords may be produced, meaning that researchers need to make decisions regarding which words are worthy of detailed focus. The chapter ends with an illustrative analysis where I revisit six of my own keyword studies, arguing that researchers should consider the benefits of giving a more reflexive account of their own decision making procedures around keywords.

1. Introduction

The synthesis of corpus linguistics and discourse analysis no longer feels in its infancy, although it could be argued that it is still a developing field. The main techniques for analysis now appear to be well-established: concordances, collocates, keywords, dispersion – although tools continue to be created and updated and consensus about the preferred algorithms or settings for such techniques continues to be in flux (see for example Gabrielatos and Marchi's (2012) paper on appropriate methods for keyness or Hardie's (2014) development of the log-ratio test). Another sign that the field (or combination of fields) is moving forward

1. The research presented in this paper was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

https://doi.org/10.1075/scl.87.04bak

can be found in the more critical, experimental and reflective research that has emerged in recent years. For example, research by Marchi and Taylor (2009, 2013) and Baker (2015) involves experiments where more than one researcher independently works on corpus data and results are then compared. Baker (2012) examines how corpus linguistic techniques have been used to identify author bias in texts (from a critical discourse analysis perspective), concluding that interpretations of such author bias, even if backed up with quantifiable information, are influenced by the position of the researcher.

This chapter continues this more reflective strand of research in corpus-based discourse analysis by focussing on the keywords technique and carrying out a piece of reflexive research on a number of keywords studies I have carried out over an eight year period. As an advocate of the keywords technique (and corpus techniques for discourse analysis in general), it is important to temper enthusiasm for one's preferred technique with a frank consideration of potential pitfalls, lest new analysts are misled into thinking they are somehow "getting it wrong" because their experience of using the technique is not congruent with how it is described by others. It is not my goal to dissuade analysts from using keywords (or corpus techniques in general), but acknowledging the issues surrounding this technique in particular should at least help to ensure against over-confidence in interpretation and reporting of results, engendering a willingness to combine the method with other techniques as a form of triangulation.

After briefly describing the keywords technique and how it relates to discourse analysis, I discuss the importance of researcher reflexivity, while the last part of the chapter involves a reflexive analysis of my own research, aimed at addressing the question of the extent that we can say that keywords offer an objective method of analysis, and indeed, if they do not – does it matter?

2. Keywords

One way of identifying the focus of a corpus, set of texts or single text is to consider the most frequent words in that data set. As Teubert (2005: 5) notes "Frequency is… an essential feature for making general claims about the discourse". In many cases frequent words tend to be from closed grammatical classes (articles, determiners, prepositions, pronouns, conjunctions etc.). Such words can sometimes be helpful in identifying author stance or positioning of certain topics (e.g. McEnery (2006) shows how the conjunction *and* was frequently used in tracts about swearing in order to create an association between swearing and other phenomena like violence and sex, while Baker and Love (2015) note that the pronoun *I* was unexpectedly infrequent in a corpus of political speeches in 2013

which were against gay marriage in the UK, suggesting the speakers wanted to distance themselves from an increasingly marginalised position). However, many frequent words are less revealing of discourse or stance, simply telling us about what is typical of language (or a particular register) in general. On the other hand, a keyword analysis focusses on words that are not necessarily the most frequent ones in a corpus, although they are more frequent than we would expect. Thus such words may otherwise be overlooked because they do not always appear at the top of frequency lists.

Briefly then, a keyword is any word "whose frequency is unusually high in comparison with some norm" (Scott 2008). The concept was developed by Mike Scott who incorporated a technique for eliciting keywords into his corpus analysis software *WordSmith Tools*. Scott goes on to say that keywords "provide a useful way to characterise a text or a genre". While a simple frequency list only requires a single corpus, a keyword list requires two corpora or sets of texts. Often one corpus is much larger than the other and acts as a 'reference' for typical word frequencies in the language (or register) under study. However, a second way of considering keywords is to compare two related texts, with each acting as the reference for the other one (e.g. a comparison of tabloid and broadsheet news stories). Traditionally, the frequencies of each word in the two corpora are compared by the software carrying out statistical tests, such as a log-likelihood or chi-square test on each word. The test also takes into account the overall size of each corpus as well as the frequency of each word.

Clearly, the comparator corpus will have an impact on what keywords emerge in the corpus under analysis, although Scott (2009) has carried out experiments by comparing different corpora, concluding that there is no such thing as a bad reference corpus. I would agree, although note that some reference corpora are better than others depending on the research questions that are set.

Keywords have certainly been an important technique in my own research. I would characterise them as "signposts", helping to point researchers towards concepts or uses of language in a corpus that they might not have realised were unusually frequent. As researchers are not usually able to read every word in a corpus, keywords can offer starting points or a "way in" to the analysis. They are reductive in that they focus our attention on potentially useful phenomena, making the analysis more manageable (and saving time). Using Tognini-Bonelli's (2001) distinction between 'corpus-based' and 'corpus-driven' approaches, we could view keywords as a corpus-driven approach – instead of commencing with a shopping list of concepts, features or words to explore, the analyst does not know where the analysis will begin, and the corpus techniques thus drive the analysis along. A reasonably convincing argument could be made then that keywords help to reduce researcher bias as computers use statistical techniques in an objective

way to provide the list of words. Especially with more critical forms of discourse analysis, our analysis needs to be rigorous and convincing, rather than resembling a one-sided polemic.

A list of keywords does not constitute an analysis. It is merely a precursor to one. There is no single way to carry out a keyword analysis, although below I have tried to systematise some of the procedures I normally carry out with a list of keywords. Although this is described in a linear way with "stages", the reality is more cyclical and recursive, with considerable movement back and forth between stages.

At an early stage we need to decide how many keywords we will examine or where our cut-off for statistical significance will be. I will discuss this in more detail later, although for the moment imagine that the decision can be easily made and we thus have a finite number of keywords to examine. We may initially want to impose some sort of categorisation scheme on these keywords, either by using a predefined system like the USAS semantic tags (Wilson & Thomas 1997) or one that is created in a more ad hoc way. It might be useful to group the keywords into grammatical categories like adjective, preposition etc. Another way could be to group words in terms of how they contribute towards certain themes or concepts – such as words that reference size or words which appear to be positive evaluations. It is recommended that quick concordance searches are carried out in order to make the classification more accurate, and that some sort of decision is made about how to categorise words which fall into more than one category (either by putting them in their most frequent category or in both or using some other consistently applied procedure). As a process of categorising words, often the categorisation scheme needs to be updated and refined, and it is often difficult to avoid a 'Miscellaneous' category for words that stubbornly refuse to link to other words. Not too much time should be spent anguishing over this stage; however, as again, it is not the actual analysis but constitutes a preliminary way of making sense of the keywords and identifying potential similarities between them.

Following the initial classification, I normally choose a keyword or set of keywords to examine (again I discuss this more in the coming reflective section). Unless the keyword is a completely unfamiliar word or the analyst has absolutely no prior knowledge or experience of the corpus and the register it is from, it can be quite difficult not to start hypothesising about why such a keyword appears in the list. Such prior hypotheses ought to be acknowledged – they can be helpful later if our earlier theories about a word's function or purpose in the corpus are actually confounded – in such cases we have lit upon what Partington, Duguid and Taylor (2013: 11) call "uncovering… non-obvious meaning" which can often be the most rewarding and memorable parts of a corpus approach to discourse analysis. An analysis which tells most readers what they already know does have some worth

but there is a deeper pleasure in unearthing the unexpected that would have been difficult to discover through other means.

Now the analysis proper can begin, first involving a descriptive analysis of each keyword. If it is frequent enough it can be investigated via its collocates and the frequent clusters it appears in. Whether it is frequent or not, (expanded) concordance analyses are also carried out, sorting concordances in various ways (e.g. alphabetically one place to the left or right). If a keyword is exceptionally frequent we may only want to consider thinned sets of concordance lines or carry out further targeted concordance analyses of particularly frequent clusters or collocational combinations. It is also helpful to consider dispersion patterns, asking whether a keyword only occurs within certain texts, time periods or textual positions within the corpus, such as at the end of texts. This will help to indicate whether a keyword is due to idiosyncrasies of a small number of authors (perhaps representing a minority discourse in the corpus) or whether it is more typical of the entire corpus.

Having produced a descriptive account of the keyword we then move to further consideration of other contextual factors. Teubert (2005: 3) notes:

> Meaning is in the discourse. Once we ask what a text segment means, we will find the answer only in the discourse, in past text segments which help to interpret this segment, or in new contributions which respond to our question. Meaning does not concern the world outside the discourse. It is up to each individual to connect the text segment to their first-person experiences, i.e. to some discourse-external ideation or to the "real world". How such a connection works is outside the realm of the corpus linguist.

As such connections are not always likely to be made if one remains inside the corpus, it can become necessary for the analyst to shift gears. This could lead to stepping outside the corpus to look at the etymology or history of a keyword, or even its dictionary definitions. Interpretation may involve examining the conditions of production and reception of some of the corpus texts asking questions like what was allowed to be said, who were the texts written by and for, and how did audiences receive the texts? We may also want to consider intertextual relationships – does the keyword appear in the corpus because it occurs in other texts that may not have been included as part of the corpus collection but are still relevant to its presence?

A further stage involves a more interpretative and explanatory analysis, involving asking *why* a keyword appears. What is it being used to achieve and does the keyword help to contribute towards a particular representation, discourse or ideological position? Also, does the keyword link through or have a similar discursive function to other keywords that have been examined?

Finally, there comes what is perhaps a more optional stage, depending on researcher motivations and the extent to which the analysis aims to be critical. We may want to evaluate the discourse that the keyword contributes to, asking who benefits from the popularisation or unquestioned circulation of that discourse (and who does not benefit), and what could be done, if anything, to improve conditions for people who the discourse does not appear to favour.

3. The keyword *Moslem*

As a brief illustration I will summarise how I analysed a single keyword, which is taken from Baker, Gabrielatos and McEnery (2013) where I compared two corpora of British newspaper articles about Islam and Muslims (published between 1998 and 2009). One was a corpus of tabloid news articles from newspapers like *The Daily Mail* and *The Mirror* (43 million words) while another constituted broadsheets like *The Guardian* and *The Telegraph* (103 million words).[2] Each corpus acted as a reference for the other, and in the tabloid corpus, the keyword with the highest log likelihood score was *Moslem*.

Moslem is a variant spelling of *Muslim*. This was not the only variant spelling that appeared as a keyword in the analysis, and particularly the spellings of Arabic words like *Taliban* and *Al Quaeda* were often spelled inconsistently, resulting in them being key too. I hypothesised that collectively, these different spellings reflected a lack of familiarity in the British press with these terms, especially during the earlier news articles prior to the 9.11 terrorist attacks. Table 1 shows a random sample of 18 concordance lines of *Moslem* taken from both corpora.

Table 1. Concordance sample of *Moslem*

beyond bare hills, a police post and a small	Moslem	shrine, until a voice from below shouts
launch an attack to mark Eid al-Adha the	Moslem	"festival of sacrifice" which ends on Saturday
Albanians, Greeks, Poles and Kurds. Turkey is a	Moslem	secular state which sees itself as very
Cleric supports 'jihad': A	Moslem	cleric fighting deportation for allegedly
growing fears they could have been helped by	Moslem	terrorists who have been flooding into
two main political parties, the Pakistan	Moslem	League led by Nawaz Sharif, the ousted
needed the money, it took an audience with a	Moslem	cleric to persuade Sukur to go to Italy
the Begs who under Turkish pressure turned	Moslem	and thus preserved their estates until
time a duel between the Catholic with his	Moslem	second against the Orthodox Serb

2. Tabloid newspapers generally do not have as many pages and contain shorter articles than broadsheets, hence the discrepancy in size between the two corpora.

Jackson had hired a security team from black	Moslem	extremist group Nation of Islam. He
where Russian forces have been fighting	Moslem	rebels. Pakistans strike at IMF tax
famously modest in his tastes; a dutiful	Moslem	, tall and articulate. He grew up as a
But she was never in any serious trouble. "In	Moslem	countries, I would refer to a husband or
was brought up as a Christian, became a	Moslem	, then converted back about 12 years ago
God's Own Game' not having made it yet to	Moslem	Lebanon* A good impression will help
Cup competition, contested by teams of	Moslem	schoolchildren in Sydney. He scored 10
ended 15 days later in Algiers. Dec 1994:	Moslem	fundamentalist gunmen seized an Air France
official, Erzimanov, did not drink as he is a	Moslem	. But you have to understand our situation

Examining these concordance lines, *Moslem* appears to be used as both a noun or an adjective, and in a range of different contexts, some which frame Moslems as violent (e.g. the references to *jihad*, *fighting*, *terrorists*, *extremist group*, *rebels* and *fundamentalist gunmen*). However, we cannot tell from this concordance whether *Moslem* is used in a way which is different from the usual *Muslim* spelling, so the collocates of *Moslem* and *Muslim* were compared to see if the words were used in different contexts (see Table 2).[3]

Table 2. Comparing collocates of *Muslim* and *Moslem*

Term	Collocates
Muslim AND *Moslem*	*community, Council, cleric, extremists*
Muslim only	*Britain, world, leaders, women, Christian, countries, communities, British, population, country, who, Jewish, a, groups, woman, men*
Moslem only	*devout, fundamentalists, Hindu, Ramadan, holy, strict, MP, Sikh, extremist, predominantly, fundamentalist, Moslems, Sarwar, Brotherhood, Shi-ite*

Some of these collocates were shared by both words, referring to general contexts like *Muslim community* or *Moslem cleric*. Words which only collocated with *Muslim* also tended to be quite general, relating to large identity groups like *British* and *women* as well as other grouping words like *groups* and *population*. However, many of the collocates of *Moslem* only were related to strength of belief words – mainly indicating strong beliefs: *devout*, *fundamentalists*, *strict*, *extremist* (although *extremists* collocates with both words). So there is some indication that *Moslem* holds a different discourse prosody than *Muslim* – associated more with strong belief. However, so far the analysis does not explain *why* this would be the case. A clue was found, however by considering patterns of dispersion.

3. Collocates were calculated via *Sketch Engine*, using a span of 5 words either side of the target word. Only the top 20 collocates were considered for each word (using the logDice statistic).

When examining the concordance lines I had noticed that the texts which contained citations of *Moslem* tended to be restricted to quite a small number of newspapers and it appeared that two tabloid newspapers in particular (*The Mail* and *The Express*) were using the word *Moslem* very frequently. I decided to examine dispersion patterns for *Moslem* more systematically, taking into account change over time and focussing on these two newspapers.

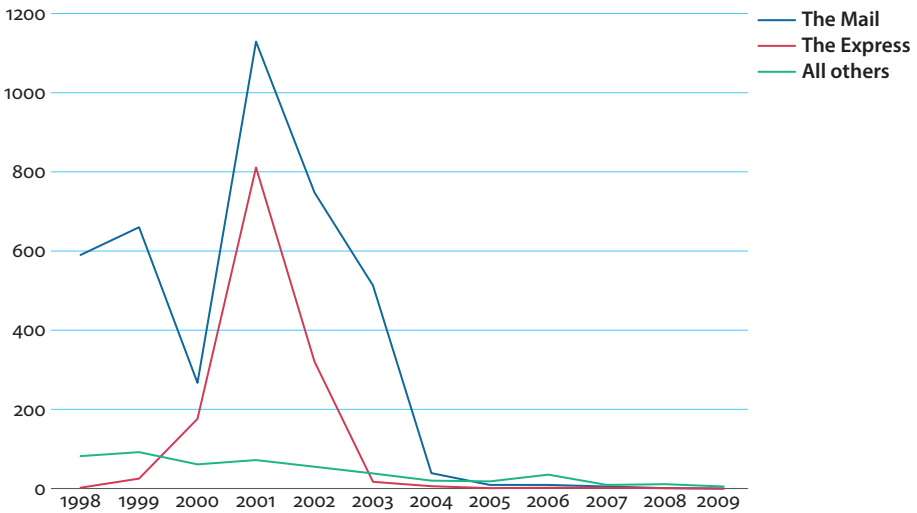

Figure 1. Change in use of *Moslem* over time (frequency per million words)

Figure 1 indicates that *Moslem* was mostly used by *The Mail* and *The Express*. All of the other newspapers considered collectively, hardly used that spelling at all. *The Mail* and *The Express* both peaked using the *Moslem* spelling in 2001, but then they too had largely dropped that spelling by 2009. However, *The Express* appears to have made a decision to reduce use of the word earlier, so by 2003 it had almost completely abandoned that spelling. On the other hand, *The Mail* continued to use the word in 2003, with its usage falling to nearly zero in 2004. The patterns in Figure 1 thus raise further questions about the motivations of certain newspapers or writers.

The consideration of other sources of information became valuable in explaining the dispersion pattern. Google searches on the terms *Moslem*, *Daily Mail* and *Daily Express* led to a link to the Muslim Council of Britain's website. The Muslim Council of Britain is an umbrella group which represents and unites Muslims across the UK. On its website an article from one of their newsletters was found, of which part is quoted below.[4]

4. The link to this excerpt was <http://www.muslimnews.co.uk/paper/index.php?article=979>. However, the article no longer appears at the site.

> Several of our Muslim correspondents were unhappy with the Daily Mail and the Daily Express continued use of the antiquated spelling *Moslem* in their stories. We wrote to the editors of both newspapers in July and urged them not to unnecessarily antagonise their many thousands of Muslim readers in this way. The Editor of the Daily Express, Chris Williams, agreed with the contents of our letter and actioned the change right away. (27 September 2002)

A link to another website indicated that the *Moslem* spelling variant was similar to the Arabic pronunciation for a word which means 'oppressor'. It is notable that *The Express* had largely stopped using this spelling after the point in 2002 when the Muslim Council contacted the newspaper (there were only 17 cases of its use in 2003), but *The Daily Mail* appears to have continued to use the spelling for a further year (513 cases of its use in 2003), until finally they appear to have joined the consensus.

The keyword *Moslem* thus reveals something interesting in terms of ideologies of different newspapers. While we know that the Muslim Council asked *The Mail* not to use *Moslem* in 2002, and while we do not have a record of how they responded, we could indirectly conclude that their continued use of the word in 2003 indicates a somewhat hostile stance. However, an important point is that the corpus analysis alone would not have told us the most pertinent information. It would have revealed only a partial account of the word's contribution to discourses around Islam in the British press.

4. Reflexivity and keywords

Clearly, supplementing keyword analysis with other approaches which take into account context is valuable. However, I want to spend the remainder of this chapter discussing the extent to which a keyword analysis actually does reduce bias. I have argued that while corpus analysis reduces human cognitive biases, we cannot remove them completely as analysts can be selective in terms of the parts of the research they report on (Baker 2006: 12).

Since the 1990s social scientists have problematised the idea of the neutral, objective researcher. Eichler (1991: 13) has written that science is not value-free while Code (1991: 35) notes "that knowledge is a construct that bears the marks of its constructors". As researchers we need to be aware that our own identities and experiences will impact on how we carry out the analysis and the discoveries we make. Baxter (2003: 587–61) thus advises that part of the research process needs to involve a reflexive analysis of how the researcher has influenced the research. With corpus approaches to discourse analysis I believe such reflexivity is especially important, particularly as the field is still relatively young, and also because we can

fall into a trap of assuming that because computers are accurate and "unbiased", they will produce an objective and fair analysis, especially if we rely on corpus-driven methods.

My reflexive analysis began when I considered a recent keywords study I had carried out. I reflected on the way I had approached the analysis, asking whether the method I had used had been clear enough. In other words, had I described how I had created the keyword list in enough detail so that someone else could replicate the study? I also wanted to see if I had provided a clear rationale or set of reasons for my decisions, and ultimately for the keywords I'd chosen to look at. A final question I considered was the extent to which my own biases drove the analysis along.

However, I decided to expand the reflection to consider a wide range of studies over time, particularly as this would help me to ascertain the extent that I had been consistent in my methodological choices over recent years. I could have expanded the study further to look at other people's research but it would have been difficult to gain the same sorts of insights into the thought processes of others, and limiting the subject of study to myself, I was at least able to look back over my earlier research and recall fairly well why I made certain decisions.

As an initial point in this reflexive process it is worth noting that the process of deriving keywords involves different kinds of funnelling or reduction. First a set of statistical funnels are applied. Analysts impose their own cut-off points, for example deciding a maximum p-value for statistical significance or choosing to only look at the first 20 or 100 keywords in the list, when it is ordered by keyness score. Additional keywords may be utilised, such as just considering keywords if they have a certain minimum frequency or occur in a certain proportion of texts in the corpus overall. Beginning users of tools like *WordSmith* or *AntConc*[5] may not realise that the number of keywords they are given are dependent on the default cut-off settings of the tool and can be altered at will if different quantities of keywords are required. The default settings of corpus tools may be other people's cut-off points but they do not offer a free pass to objectivity.

Secondly, once a list of keywords is obtained, there is often another form of funnelling which is more intuitive. Certain keywords are selected for a more detailed analysis than others, while some keywords may be overlooked completely. Table 3 indicates six studies which I had authored or co-authored where I had used keywords analyses.

5. Software tools by Mike Scott <http://www.lexically.net/wordsmith/> and Laurence Anthony <http://www.laurenceanthony.net/software.html>.

Table 3. Summary of keywords studies

Study	Corpus tokens	Funnelling 1		Funnelling 2
		p-value	Keywords elicited	Keywords reported (%)
Baker (2005) Chapter 2	100,000	0.0005	41	27 (66%)
Baker (2005)* Chapter 6	2 million	0.000001	1055	197 (19%)
Baker (2006)	130,000	0.000001	22	19 (86%)
Gabrielatos & Baker (2008)	140 million	0.00000000000001	1,500+	250 (16%)
Baker (2010)	87 million	0.000001	Not said but top 300 considered	172 (57%)
Baker, Gabrielatos & McEnery (2013)	143 million	Top 100 in 11 lists considered	717 unique keywords	114 or 19 in detail (16% or 0.3%)

Note: Baker (2005) contains two separate keywords studies which are described in Chapters 2 and 6 of that book.

The second column gives the size of the corpus in terms of tokens involved. So for example, Baker (2005) Chapter 2 in the first row is my study of the House of Lords debates on the age of sexual consent. I compared speech of people who wanted to change the age of consent with those who wanted to keep it the same. But collectively, all the speech came to 100,000 words. The next two columns indicate the statistical funnelling involved. I have noted the p-value that was employed in obtaining keywords, and how many keywords this p-value produced. For the last study, a p-value was not specified but instead we considered the 100 keywords that had the highest log likelihood scores (across 11 keyword lists). Also, in Baker (2010) the p-value was specified but I did not indicate how many keywords this produced in total. Instead I looked at the top 300 keywords.

From these columns it can be seen that the size of the corpus appears to have an effect on the number of keywords produced. Corpora of around 100,000 words can produce relatively small numbers of keywords, especially if *WordSmith*'s default p-values are used (e.g. Baker 2006). When working with corpora of many millions, we are more likely to obtain hundreds or even thousands of keywords. So Baker (2005) Chapter 6 which used a 2 million word corpus, produced over 1,000 keywords at *WordSmith*'s default setting. Thus, the larger the corpus, it appears that the more analytical work is created, unless the p-value is made smaller.

The final column, which is perhaps the most pertinent, asks how many keywords were actually reported on in the analysis. Both the actual number and that number as a percentage of the number of keywords elicited are noted. For example, Baker (2005) Chapter 2 found 41 keywords, although 27 of them were

actually referred to, or 66%. Notably, none of the six studies referred to every keyword elicited. Even the two smallest-scale studies resulted in reports of 66% and 86% of the keywords that had been available. With some of the larger studies, a much smaller percentage of keywords were reported on. Baker, Gabrielatos and McEnery (2013) derived 717 keywords of which only 114 (16%) were mentioned and of those only 19 (0.3%) were actually discussed in any detail.

However, this last point: "discussed in any detail" raises a further question about what an analysis and its write-up actually involves. So I returned to the studies to consider the level of detail that I went into with the description of these keywords. In some cases a relatively large amount of space was devoted to a single keyword. For example, there was a 1,000 word analysis of the word *criminal* in Baker (2006) which also included a screenshot of its collocates and two concordance tables. Yet considering that many journal articles and book chapters are between 7,000–12,000 words (with the analysis sections usually comprising about half to three quarters of that) it is unlikely that 1,000 words can be devoted to more than a handful of keywords. A more succinct account was given to the keywords *Osama*, *Bin* and *Laden* in Baker (2010: 326) which are covered jointly within the space of 125 words. However, many of the keywords I have examined are dispatched even more briefly. In the same study the keyword *community* is given 22 words: "Additionally, the word *community* occurs as a broadsheet keyword with its most common collocates being *Muslim*, *British*, *large*, *support*, *relations* and *backlash*" (Baker 2010: 329). Finally, some words are only peripherally encountered, as with the keyword *cronies* in Gabrielatos and Baker (2008). This word only ever appears in a table, categorised as a type of threat and then sub-classified as crime. Essentially then, there are two words devoted to its analysis: *threat* and *crime*.

It should be clear then that a typical keywords analysis (at least the way I approach it) involves subjective decisions at numerous levels. It is very likely that had the same corpus been given to a different set of researchers, they may have selected different cut-off points for statistical significance and then chosen different keywords to analyse from the resulting lists that were elicited. They would also have allotted varying amounts of space to the analysis of individual keywords.

Of course, it is not always possible to apply a consistent set of standards across multiple studies. Each project is different, containing corpora of various sizes and with unique research questions, timeframes and publishing goals. Additionally, corpus studies of discourse are developing so it is not always wise to hold on to older techniques if new ones are demonstrated to be more useful. Having uncovered quantitative inconsistency across my own studies, it was next decided to take a more qualitative approach where I tried to reflect on why I had chosen to consider certain keywords for analysis and/or inclusion in their subsequent published research paper while others had been overlooked.

I will limit this part of the reflexive exercise to the last study in Table 3 which considered keywords in a set of 11 newspapers in articles about Islam. This was the most recently completed study so I was most able to recall the thought processes that resulted in certain decisions. I reread the analysis chapter and notes I had made, along with the full tables of keywords that had been produced. I recalled why I had been drawn to certain keywords and others were passed over. In this study, there were a total of 717 elicited keywords with the analysis limited to a 4,000 word section of a book chapter. The analysis actually referred to 114 keywords although many of these were mentioned "in passing" and only 19 were discussed in more detail (receiving more than a couple of sentences). How had I decided what to focus on?

First, looking back at the keyword lists, I had given precedence to the strongest keywords or those with the highest log likelihood scores. Second, some keywords were familiar to me and triggered hypotheses about why they appeared in certain newspapers. Their analysis provided a good demonstration of the ideology of that newspaper. For example, the word *hook* appeared as a keyword in the tabloid newspaper *The Sun*. I hypothesised that this word referred to a Muslim preacher called Abu Hamza who had a hook for a hand. The analysis of *hook* confirmed this and also provided cases of the newspaper's focus on this particular Muslim who was constructed as a pantomime villain with headlines that made puns about the hook including: "our soft lawyers let evil Hamza off the hook", "200 thousand pound right hook, Hamza's lawyer hits YOU for massive legal aid bill' and "Sling your hook!".

Alternatively, some words were analysed because I could not hypothesise a reason for their appearance in the keyword lists. For example, words relating to higher-education degrees (*MBA*, *MPhil*, *MA*, *MSc*) were key in *The Guardian*. I thought it was unlikely that the newspaper would be referring to Muslims with degrees and the subsequent analysis of these words revealed that *The Guardian* published supplements that list postgraduate courses across the UK, including courses like 'Contemporary Islamic Studies' and 'Islamic Political Economy'. Additionally, keywords in *The Observer* were *toponymic* and *patronymic*; both words were unfamiliar to me and so I could not speculate any reasons for their presence in articles about Islam. Concordance analyses revealed that they occurred in a very long article on the origins of British surnames which briefly mentioned surnames that came from "the Muslim world". As these words were poorly dispersed across *The Observer* and barely related to the topics of Muslims and Islam, I only devoted a couple of lines to them.

This leads to a fourth motivation for considering a keyword in more depth, which was the extent that I felt the word could provide a good answer to the research question I wanted to ask – in this case, what distinguished each newspaper in terms of how it discursively represented Muslims and Islam? Therefore, words like *Islamic* (*The Telegraph*), *Islamist* (*The Guardian*) and *sharia* (*The Express*)

were felt to be particularly relevant, as were words relating to strength of religious belief, e.g. *zealotry* (*The Times*) and *extremists* (*The Express*). However, at a first glance other keywords did not appear especially relevant to the research question. For example, *Times* keywords included *horticulture*, *introductions* and *wholesaler* which I did not consider to be particularly relevant and I confess I did not analyse.

Examining concordances of these words now, *horticulture* sometimes appears in articles which list awards given for various services, of which Muslims are mentioned elsewhere. The keyword *introductions* does not appear to have much connection to Islam or Muslims but appears in a range of different contexts, particularly detailing narratives involving 'society' where formal introductions are made (this may be linked to the newspaper's assumptions about the social class of its readers and their interests). Finally, *wholesaler* relates to factual articles about food of which a few articles involve Asian or Muslim business-people. While such keywords do not tell us much about explicit representations of Muslims, it was perhaps presumptive to completely ignore them as "in passing" representations that are not connected to negative stereotypes around Islam: terrorism, extremism, war etc. are worth focussing on, in order to give a sense of balance.

Another related set of overlooked keywords appeared to be related to the writing style or topic focus of the newspaper, rather than telling us much that was specifically going to answer the research question. For example, the keyword *stunning* in *The People* (a Sunday tabloid) was almost always used to refer to female models and celebrities or to views from holiday resorts and reflected that this newspaper tended to print a large number of articles from the genres of entertainment and leisure. On the other hand, some "stylistic" keywords actually provided unexpected cases of intertextual referencing. For example, keywords in the *Daily Star* (a tabloid) included several non-standard forms like *aint*, *cant*, *cos*, *goin*, *gud*, *hav*, *luv*, *ov*, *pls*, *ppl*, *shud*, *ur*, *wen*, *wiv*, *wot*, *wud* and *ya*. Initially I wondered if this was due to an exceptionally informal journalistic style, but concordance analysis revealed that they appeared in a regular column called *Text Maniacs* where readers' text messages were published; hence the use of forms often associated with computer mediated communication. Such text messages often commented on articles that had been published the previous day and it was interesting to compare how the original article was written and then see how various readers interpreted the article. For example, consider the following short article in *The Star* (October 24, 2005):

> PIGGYBANKS are facing the axe – because some Muslims could take offence. Britain's top High Street banks have ruled the money-boxes are politically incorrect. But last night the move sparked snoutrage. And one of Britain's four Muslim MPs, Khalid Mahmoud, said: "A piggybank is just an ornament. Muslims would never be seriously offended." (*The Star*, October 24, 2005)

The following day, a Text Maniac used the keyword *ppl* in the following message (translated into standard English underneath):

> muslims r offended by our piggy banks! ? Then the £56 me n ma wife n ma 4 girls have got in our piggy bank 2 help the ppl in pakistan wil b spent on a fry up.
>
> (Muslims are offended by our piggy banks! ? Then the £56 me and my wife and my 4 girls have got in our piggy bank to help the people in Pakistan will be spent on a fry up)

Despite the fact that the article did not actually quote any Muslims as being offended, and even quoted from a Muslim MP who was not offended, the reader of the article who sent the text messages appears to focus on Muslims who *are* offended and then makes the provocative statement that rather than giving money to a charity to help people in Pakistan, they will instead have a "fry up" (which normally involves consumption of pork products which are forbidden in Islam). It is possible that people who did not read the original article but only saw the text messages about the following day would get the impression that Muslims *were* offended by piggy banks, a point which was the opposite of the point made in the initial story. In such ways, reader's intertextual contributions can provide additional layers of meaning to representations of Muslims.

One self-imposed criterion involved whether a particular keyword appeared to have a similar function or meaning to other keywords in a list. As described earlier in this chapter, I had conducted an initial categorisation of keywords, putting them into groups such as pronouns, reporting words, words referring to the media etc. To ensure as much coverage as possible, it could be a good idea to try to consider at least a couple of words from each category. One of the categories of words which emerged was proper nouns, and quite a lot of these referred to names of journalists who wrote for individual newspapers. While it was useful at later points in the analysis to discuss certain journalists who wrote about Islam in certain ways, it was felt that discussing journalists in the chapter would have resulted in a repetitive analysis, so these words were not given space. One proper noun which was unexpectedly useful, however, was the word *Faisal*. This keyword occurred in *The Observer* and concordance analyses showed that the word almost always appeared in the 'Business' supplement of that newspaper. However, upon reading such articles there appeared to be no references to Muslims or Islam within them. The mystery was solved when it was discovered that the articles were all written by the same author: Faisal Islam. When the corpus had been collected from the database *Nexis*, we had stipulated a set of search terms that articles should contain, and one of these had naturally been the word *Islam*. We had provided a list of exceptional cases to avoid irrelevant hits, but we had not envisaged that

Islam would be frequently used as a surname and thus result in a set of articles that were not wanted. The presence of the keyword *Faisal* thus showed a (relatively minor) flaw in the compilation process of the corpus, illustrating an unexpected benefit of a keywords analysis.

The process of returning to the decisions that I had made regarding which keywords to look at raised a number of questions about whether I did justice to the analysis. Considering the externally imposed word count restrictions, it is doubtful that I could have written about many more keywords without losing some other part of the book, but a question arises as to whether I focussed on the right keywords or whether the decision-making processes I engaged in were rigorous enough and also replicable. As I did not spend much time describing the process, I doubt the latter would be the case. And for the analysis to be truly rigorous, all 717 keywords ought to have been subjected to the sorts of enquiry detailed in the earlier part of this chapter, even if such keywords did not make their way into the chapter. At least in terms of transparency, all 717 keywords were articulated in tables in the chapter, although their lack of discussion would leave it up to the reader to guess at why they appeared and how relevant they were.

5. Conclusion

Keywords often signify the commencement of a great research journey, pointing us in directions we would never have thought to take. As an example, my discussion of the keyword *Moslem* enabled me to uncover how this spelling variant was characteristic of two newspapers in the UK press, who appeared to react differently when asked to stop using it. The corpus data itself was not able to provide the "story" behind the word, but I would not have known that a story existed, had the word not appeared as a top keyword initially.

Thus keywords can result in a very productive and targeted form of analysis, although we should be aware that they can elicit so many potentially interesting words for analysis that it is difficult to do justice to them all. If keywords are signposts, then it is still the responsibility of researchers to decide which ones they will follow. As a result of the reflexive exercise I engaged in for this chapter, I suspect I have taken certain routes that often led me to destinations that were familiar, provided "attractive" findings or were easier to get to.

I can only confidently speak for myself, but having supervised numerous dissertations and reviewed many journal articles where keyword analyses are involved, I suspect that the practices I engage in around keyword analyses are not unusual. It is important to bear in mind then that keywords help to reduce bias but they do not remove it. In a worst case scenario they *obscure* bias. As a way forward

then, it is recommended that researchers attempt to incorporate greater reflexivity into their corpus analyses of discourse, attempting to question why certain words were given precedence over others, and being transparent about the limitations of this technique and others like it.

References

Baker, P. 2005. *Public Discourses of Gay Men*. London: Routledge.

Baker, P. 2006. *Using Corpora to Analyse Discourse*. London: Continuum.

Baker, P. 2010. Representations of Islam in British broadsheet and tabloid newspapers 1999–2005. *Language and Politics* 9(2): 310–338. https://doi.org/10.1075/jlp.9.2.07bak

Baker, P. 2012. Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies* 9(3): 247–256. https://doi.org/10.1080/17405904.2012.688297

Baker, P. 2015. Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis. In *Grammar, Text and Discourse: In Honour of Susan Hunston*, M. Charles, N. Groom & S. John (eds), 283–300. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.73.13bak

Baker, P., Gabrielatos, G. & McEnery, T. 2013. *Discourse Analysis and Media Bias: The Representation of Islam in the British Press*. Cambridge: CUP.

Baker, P. & Love, R. 2015. The hate that dare not speak its name? *Journal of Language, Aggression and Conflict* 2(2): 57–86.

Baxter, J. 2003. *Positioning Gender in Discourse: A Feminist Methodology*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9780230501263

Code, L. 1991. *What Can She Know?: Feminist Theory and the Construction of Knowledge*. Ithica NY: Cornell University Press.

Eichler, M. 1991. *Non-sexist Research Methods: A Practical Guide*. London: Routledge.

Gabrielatos, C. & Baker, P. 2008. Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005. *Journal of English Linguistics* 36(1): 5–38. https://doi.org/10.1177/0075424207311247

Gabrielatos, C. & Marchi, A. 2012. Keyness: Appropriate metrics and practical issues. Paper presented at CADS International Conference 2012, 13–14 September, University of Bologna Italy.

Hardie, A. 2014. Statistical identification of keywords, lockwords and collocations as a two step procedure. Paper presented at ICAME 35 Conference 30 April – 4 May, University of Nottingham.

Marchi, A. & Taylor, C. 2009. If on a winter's night two researchers… A challenge to assumptions of soundness of interpretation. *Critical Approaches to Discourse Analysis Across Disciplines* 3(1): 1–20.

Marchi, A. & Taylor, C. 2013. Experimenting with objectivity in corpus and discourse studies: Expectations about LGBT discourse and a game of mutual falsification and reflexivity. Paper presented at Corpus Linguistics Conference July 23, 2013, Lancaster University.

McEnery, T. 2006. *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

Partington, A., Duguid, A. & Taylor, C. 2013. *Patterns and Meanings in Discourse. Theory and Practice in Corpus-assisted Discourse Studies (CADS)* [Studies in Corpus Linguistics 55]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.55

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

Scott, M. 2009. In search of a bad reference corpus. In *What's in a Word List?*, D. Archer (ed.), 79–92. London: Ashgate.

Teubert, W. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13. https://doi.org/10.1075/ijcl.10.1.01teu

Toginini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Wilson, A. & Thomas, J. 1997. Semantic annotation. In *Corpus Annotation: Linguistic Information from Computer Texts*, R. Garside, G. Leech & A. McEnery (eds), 55–65. London: Longman.

Europhobes and Europhiles, Eurospats and Eurojibes

Revisiting Britain's EU debate, 2000–2016

Alan Partington and Matilde Zuccato
University of Bologna

This paper is an examination in three parts of the UK's debate on membership of the European Union, before and immediately after the so-called 'Brexit' Referendum. The first part takes its cue from an article by Wolfgang Teubert which has exercised considerable influence in the field of corpus-assisted discourse studies (CADS), namely, his examination of the language of EU-scepticism in the UK (Teubert 2001). The aim of the CADS approach is the uncovering, in the discourse type under study, of *non-obvious* meanings and patterns of meanings, that is, meanings which might not be readily available to naked-eye perusal. Teubert's paper was an inspiring example of these procedures.[1] In the second part, a para-replication of Teubert's work, we revisit attitudes to the EU as represented in sections of the UK press in 2013 (the year the British Prime Minister announced an "in-out" referendum on EU membership). The third section examines the themes debated immediately before and immediately after the referendum vote in June 2016. We also reflect on how the much-invoked notion of *negative representation* needs to be employed with care, particularly with regard to media discourses.

1. Introduction: Teubert's 2001 study on Euroscepticism

In this Section, we revisit Teubert's celebrated study of EU-sceptic attitudes in the UK in 2000 (Teubert 2001) and then compare these to attitudes to the EU as expressed in one left-leaning and one right-leaning UK newspaper in 2013, the year in which the UK Prime Minister declared an intention to hold a referendum

1. The paper is one of those cited in Wikipedia's entry on 'Corpus-assisted Discourse Studies': <http://en.wikipedia.org/wiki/Corpus-assisted_discourse_studies>, as well as in the *Encyclopedia of Applied Linguistics* entry on 'Corpora and Political Language' (Partington 2012).

https://doi.org/10.1075/scl.87.05par

on Britain's continued membership of the EU. The aim was to ascertain whether EU-sceptic attitudes had changed, but also to examine the pro-EU voices.

Teubert's was not the first corpus-assisted study of discourses on the EU; that honour goes to Mautner's (1995) seminal article *Only Connect*. Mautner contrasts the representations of EU issues and news actors emanating from news sources from opposite sides of the political spectrum, in her case the right-leaning *Sun* and *Telegraph* and the left-leaning *Mirror* and the *Guardian*. In similar contrastive fashion, in the present work, we also examine the discourses in right-and left-leaning UK newspapers.

Teubert himself leaves us in no doubt where his heart lies. He is openly pro-EU and pro-integrationist. The paper's conclusion contains a moving picture of what a confident, visionary EU (he calls it "Europe") should be like: "a cradle of civil liberties and human rights", "a multicultural and multilingual Europe" where all citizens "engage in a polyphonic, pluralistic and multicultural dialogue" and where governments "protect the rights of free citizens rather than building increasingly sophisticated surveillance systems" (Teubert 2001: 27).[2] He senses, however, that the EU has fallen short of these ideals and part of the reason why are the discourses of fear and mistrust which emanate from EU-sceptics, those who do not share or trust these ideals.

Teubert's methodology is as follows. He collected a corpus of approximately 1,000 texts taken from EU-sceptic websites during February and March 2000. He lists these sites in order to provide the opportunity for others to replicate his study. Most of the texts included in Teubert's corpus were originally from newspapers since, as he points out, "the British eurosceptic discourse can best be observed in newspaper texts" (Teubert 2001: 24), as indeed can that of their antagonists, and the press is the arena in which the voices interact.

Analysing word-frequency lists deriving from these texts, he identifies a number of key themes in the discourse which he represents by concordances of the following items:

> *federal*, *superstate*, *bureaucra**, *unelected*, *faceless*, (*un*)*accountable*, *monster*, *Frankfurt*, *dictat**, *submerg**/*subsum**, *altar*, *sacrifice*, *worship*, *corrupt*, *corruption*, *province*, *Anglo-Saxon*, *value*(*s*), *sovereign**, *prosperity*, *independen**, *Commonwealth*, *transatlantic*, *NATO*, *community*, *global*.

We might attempt to group some of these items into categories, psychological as much as political, that is, categories by anxieties, as in Table 1.

2. The page numbers refer to the version available online.

Table 1. Various EU-sceptic anxieties according to key vocabulary items from EU-sceptic websites in 2000

Anxiety	Vocabulary
Loss of democracy	*unelected*, *faceless*, (*un*)*accountabl*, dictat**
Loss of sovereignty	*sovereign**, *independen**
Loss of UK's importance and "size"	*federal*, *superstate*, *monster*, *submerg**/*subsum**, *province*
Loss of Anglophone connections	*Commonwealth, transatlantic, NATO, Anglo-Saxon, community, global*
Corruption	*corrupt*/*corruption*
Economics	*Prosperity*
EU's "messianic" ideological drive	*altar*, *sacrific**, *worship*

Most prominent is the anxiety of loss, but this connects with several other worries, for example, over size and over distance. Some EU-sceptics in 2000 saw the UK faced with an exaggerated dichotomy, between a dark future entirely enveloped within the European Union and entirely cut off from the rest of the world, and a brighter future of global, or at least Anglophone, belonging. Choosing the EU option would entail a loss of special Anglophone values and identity. Chief among these values to disappear would be "Anglo-Saxon" style democracy, a supposedly superior version to that practiced in the EU, most of whose members have little historical experience of democracy. Another loss is of sovereignty, viewed as being *taken away by* rather than *pooled with* other EU nations. Finally, there is loss of importance: the UK will be reduced in political size, become a mere "province", with real power residing far away in *Brussels* (surprisingly missing from Teubert's list) and *Frankfurt*. There seems to be no gain in choosing the EU future in this scenario, except of course in layers of *bureaucracy*.

Teubert's picture of (extreme) EU-sceptic discourse was, of course, a composite – by definition, since it is derived from analysis of a corpus, a collection of texts from various origins. It belonged to a community and did not necessarily correspond exactly to the views of any particular individual or indeed individual newspaper. And Teubert adds a singularly important caveat, to be etched in memory by all text linguists (of which corpus-assisted discourse analysts are a sub-category), that "we cannot even be sure that the writers of these texts believe in what they write" (Teubert 2001: 25), to which we might add a second qualification that individuals – and newspapers – do not always believe or express all the same things all of the time.

The next Section contains an outline of a corpus-assisted study of recent attitudes to the EU as represented in parts of the UK press, both to track their

evolution since Teubert's 2001 study and also to see how precisely they differ in relations to the left-right political spectrum.

2. The para-replication of Teubert's study thirteen years later – 2013

For this para-replication study (para-replication is replication using a different data-set (Stubbs 2001: 124, Partington 2009: 293–294)), two corpora were compiled, parallel in structure, of all the articles published in the first nine months of 2013, following the announcement by the UK Prime Minister Cameron in January 2013, from the right-leaning *Daily Mail* and from the left-leaning *Guardian*, named respectively 'DM13_full' and 'GN13_full'. The corpora were compiled by downloading each day's edition of the newspapers from the *LexisNexis* database. Both papers are influential in British political life and their web editions are the most accessed of all UK newspaper websites. As Teubert (2001) himself observes, the former is staunchly EU-sceptic, whilst the latter is generally reputed to be pro-EU. Another two, more focussed, corpora were compiled using *LexisNexis*, each one containing all the articles in the two newspapers whose headline or leading paragraph contained the items *eu* OR *european union* OR *brussels* OR *frankfurt* NOT *sport*[3] for the years 2013, named respectively 'DM13_EU' and 'GN13_EU'. Another similar pair of focussed corpora were compiled from 2005, named 'DM05_EU' and 'GN05_EU' in order to conduct a diachronic comparison to examine whether the newspapers' stances to the EU have altered as a result of changing circumstances, especially the Eurozone crisis. Most of the analyses reported below were conducted on these focussed corpora but it was sometimes helpful to check information in the larger 'DM13_full' and 'GN13_full' corpora.

DM13_full contains 17,287,000 word tokens and GN13_full contains 23,587,000 word tokens; DM13_EU contains 226,000 and GN13_EU contains 436,000 (see Table 2). This means that the *Daily Mail* EU corpus is 1.3% of the entire *Mail* 2013 corpus, whilst the *Guardian* EU corpus makes up 1.85% of the entire *Guardian* 2013 corpus.

The *European Union* and the *EU* is mentioned 2,649 times in DM13_full, equivalent to 153 occurrences per million words (pmw) and 4.157 times in GN13_full, equivalent to 176 times pmw.

3. Including this term in the *LexisNexis* search string excludes articles containing *Brussels* and/or *Frankfurt* which relate to sports events and not politics (of which there were many).

Table 2. The corpora used in the study of attitudes to the EU represented in the *Daily Mail* and the *Guardian* newspapers in 2005 and 2013 (the study began in October 2013)

Name	Contents	Size
DM13_full	All articles from the *Daily Mail*, Jan – Oct 2013	17,287,000
GN13_full	All articles from the *Guardian*, Jan – Oct 2013	23,587,000
DM13_EU	*Daily Mail* articles mentioning the EU, Jan – Oct 2013	226,000
GN13_EU	*Guardian* articles mentioning the EU, Jan – Oct 2013	436,000
DM05_EU	*Daily Mail* articles mentioning the EU, Jan – Dec 2005	333,000
GN05_EU	*Guardian* articles mentioning the EU, Jan – Dec 2005	715,000

A para-replication was conducted of an examination (Dugalès & Tucker 2012) of the proportion of mentions of the three main organs of the EU. The search terms used were as follows: for the Commission *european commission*/*EU commission*; for the Council *european council*/*EU council*/*council of ministers*; for the Parliament *european parliament*/*EU parliament*. The results can be seen in Table 3. They corroborate the earlier findings that the Commission is by far the most newsworthy of the organs. One change is that reporting of the Council has fallen still lower whilst that of the Parliament has risen, relative to the other organs (possibly as a result of the greater powers accorded to the Parliament in the Lisbon Treaty of 2009).

Table 3. The number of occurrence and percentages of mention of the European Parliament, Council and Commission in the GN13_EU and the DM13_EU corpora

	Parliament	Council	Commission	Total
Guardian	157 31.5%	36 7.2%	306 61.3%	499 100%
Daily Mail	114 33.0%	22 6.4%	209 60.6%	345 100%
Average	135.5 32.1%	29 6.9%	257.5 61.0%	422 100%

2.1 Key item analysis

Using *WordSmith Tools* (Version 5 Scott 2008), 1-, 2- and 3- word key-item lists were prepared, first of DM13_EU contrasted with GN13_EU, and then vice versa.

2.1.1 *The* Mail *key-items*

The *Mail* in 2013 has its own special term for the *Eurozone*, which it occasionally calls *euroland*. This item appears high in single-item (1-gram) keyword list, being used 52 times, whereas it never appears in the *Guardian*. On closer concordance inspection, it is sometimes used as a straight replacement for *Eurozone*:

(1) the [Italian] election outcome and rejection of all-out austerity is an enormous challenge to **euroland** stability (*Mail*, 26/02/2013)

but, in corroboration of Marchi and Taylor's (2009) observation that new coinages on *euro** are often employed for negative evaluation it is also often used critically for example,

(2) "This brutal decision by **euroland** finance ministers and the International Monetary Fund […]" (*Mail*, 18/03/2013)

(3) […] exposing the fact that **euroland**'s banking system is built on sand. (*Mail*, 18/03/2013)

(4) Osborne looks to be having some success in his fightback against **euroland** nationalism. (*Mail*, 27/09/2013)

Indeed, there is one entire article in which *Euroland* is sarcastically likened to a failed theme park:

(5) Widespread protests erupted today over the decline of the ailing **Euroland** theme park. Euroland opened in a blaze of glory, aiming to emulate the success of parks operated by Disney and Universal in the United States. But it has been plagued with problems from the start […]
Visitors complained that they were excluded from some of the most popular rides, including the European Gravy Train, which runs twice a month between Brussels and Strasbourg […]
A new attraction, called Never Never Land was built specifically for visitors from the Club Med countries, who were allowed to charge everything to Euroland credit cards […]
This one-size-fits-all currency began to unravel when it became apparent that the holidaymakers in Never Never Land were running up bills they couldn't afford to repay […]
It appears just a matter of time before the banks call time on **Euroland** and the liquidators move in. (*Mail*, 13/09/2013)

What this censorious view of *euroland* excludes of course is the fact that the UK was just as profligate in its debt-fuelled spending over the previous decade. However, the newspaper does elsewhere often criticise the over-spending of the previous Labour government (it is of course not uncommon for a newspaper to expound conflicting views on different occasions, see Section 6).

The period in question was clearly one of some elation but also trepidation for the *Mail*. The trepidation is evoked by EU plans to open borders to migrants from Romania and Bulgaria in 2014. The *Mail* key-item lists contain the following: *immigration* (129 occurrences), *immigrants* (52), *influx* (29), *migrationwatch* (a

right-leaning think-tank, 20), *eastern europe/european/europeans* (35), *Romania and Bulgaria* (32), whose citizens will obtain *free access* (9) to the UK jobs market or claim *benefits* (8). Concordancing of *migrants* and *immigrants* also revealed co-occurring items like *countless*, *large-scale*, *new wave* and *open* [Britain's] *doors* (5). Whilst it may come as no surprise that a right-leaning newspaper should be worried about mass immigration, it should be noted that anxiety over large-scale "influx" of foreigners from "undesirable" parts of Europe did not figure in Teubert's EU-sceptic discourse; the consequence of the combination of EU expansion and the free movement of people was a yet un-dreamt anti-EU nightmare in 2000.

The elation instead stems from two announcements – or *pledge/s* – made by the Prime Minister, Mr Cameron (*Mr Cameron* is itself a DM13_EU key item, with 260 occurrences) in January 2013 (see Charteris-Black (2013: 219–240) for a detailed analysis of this particular speech).[4] The first was to *renegotiate* (38) Britain's relationship/treaties with the EU to obtain a *looser, more trade-based relationship*, also described by the paper as to *claw back powers* (30) from *Europe*, or from *Brussels*. The item *Brussels* itself is a DM13_EU keyword, and although the use of metonyms rather than the official name of an institution is of course very common in political reporting, an abundance of use may be a sign of distancing.

The word *sovereign** was, we might recall, one of Teubert's keywords in Eurosceptic discourse. Whilst the item itself is actually found *less* frequently in DM13_EU than in GN13_EU, perhaps for reasons of linguistic register (the *Guardian* often uses more formal prose), it seems that EU-sceptic discourse has shifted; there is no longer a *fear* of losing sovereignty – it *has* been lost and needs, as we saw, to be *clawed back* (30). The *Guardian* too recognises the rhetorical trope of *repatriating* powers from the EU (see below).

The second of Mr Cameron's pledges is to hold, if the Conservative party wins the next elections, an *in-out* (or *in/out*) *referendum* (67) on Britain's membership of the EU, following the renegotiation of terms of membership.

A collection of other *Mail* complaints regarding the EU are signalled in the key-items lists by *taxpayers*, *British/UK taxpayers* (20), *wind farms*, **fired power stations* and *chauffeur** (17). These seemingly disparate items are linked by the theme of EU wastefulness. British taxpayers are "coughing up", "forking out", "pay[ing] for a disgraceful euro-larceny" and on schemes which "fritter away" taxpayer's cash including inefficient wind-farms replacing gas- and coal-fired power stations as well as an *army* of 500 chauffeurs and cars for *Mr MEP*:

4. Found at (both video and text): <http://www.bbc.co.uk/news/uk-politics-21013771> (15 August 2018)

(6) Safely **chauffeur**ed to the office, Mr MEP wants to send a letter to a colleague but does not have an envelope. The guide explains – in four European languages – how this problem is resolved with the help of three other people. The assistant brings him an envelope. The MEP puts the letter in the envelope and gives it to the messenger. The messenger will deliver it to the postman. (*Mail*, 29/06/2013)

These observations need however to be seen in context. If we concordance *taxpayer/s* in DM13_full (2167 occurrences, 125 occurrences pmw), we find that resentment against institutional wastefulness and official spendthrifts is certainly not confined to the EU. The *Daily Mail* sees itself as holding all officialdom – local, Westminster and EU – to financial account. Moreover, *taxpayer*/*s* is hardly an infrequent item in *Guardian* news prose; it appears 1409 times in GN13_full (55 times pmw).

2.1.2 *The* Guardian *key items*

The GN13_EU key items reflect the newspapers greater interest in international – including extra-EU – affairs relative to the *Mail*, with *Israel*, *Israeli*, *Syria*, *Syrian*, *Hamas* and *Hezbollah* high in the keyword list. As regards *Syria*, incidentally, whilst the *Mail* often calls its leader *dictator Bashar al-Assad* (24 times), the *Guardian* never does so and prefers to refer to him as *President Assad* (80 occurrences).

If the *Mail* is alternately furious with or scathing about the EU in general, there are plenty of signs in the *Guardian* key-item lists that all is far from well. Indeed, if we begin by examining the key-item lists of GN13_EU compared with the GN05_EU corpus of *Guardian* articles mentioning the EU in 2005, and vice versa, they reveal a most dramatic picture of crisis in the EU and the deterioration of UK attitudes to it. The key items of GN05_EU contains hints of "normal" gripes about the EU, about its *agricultural policy* (124–19), and also *farmers* (243–7), *rebate* (531–13), *subsidies* (388–23), *tariffs* (86–12), *quotas* (76–9), and also of environmental policy with *emissions* (116–9) and *climate* (133–31) (the first figure in each pair is the number of occurrences in the target corpus, the second the occurrences in the comparison corpus). But there are also hints of a certain optimism about the future: *enlargement* (141–4), *expansion* (70–9), *reform* (368–101), *liberalisation* (62–5).

All of these fade into the background in GN13_EU, to be replaced by *austerity* (181–0), *bailout/s* (155–0), (*euro*) *crisis* (357–260), *recession* (80–13), (*youth*) *unemployment* (192–21), *troika* (43 occurrences but this item is never mentioned in the *Mail*, see Section 6 below) and *Greek*, *Greece*, and *Greece's* (274–142). Concordancing of *Greek*, *Greece's* reveals, unsurprisingly, co-occurring items such as *debt*/*s* (23, variously described as *monumental*, *gargantuan*, *staggering*) *crisis* (20) and even on one occasion *civil war*.

There is the entirely new prospect of *repatriation of powers* (46–0), *renegotiation of terms* (of *Britain's/the UK's membership* (33–0) and even *exit* (78–10) and *leave the EU* (82–7). Interestingly, many of the items are the same as those found in the DM13_EU versus GN13_EU key-item lists, including *in-out referendum* and even *Romanians and Bulgarians*. This reminds us that key-item analysis needs to be treated with considerable caution. It is possible to obtain a radically different picture of a newspaper's stance depending on the reference database used, that is, depending on what you contrast it with. Compared with the *Mail* in 2013, the *Guardian* seems unconcerned about immigration, whereas compared with itself in 2005 it seems to pay quite some attention to it. There is also the appearance in GN13_EU of the term EU *democratic deficit* (11–0). The phrase is sometimes placed within distancing quotation marks and is often attributed to some other voice than the newspaper's. It is at times discussed, and partly dismissed, as a perception problem:

(7) Three years after the crisis began, Euroscepticism and worries about the **democratic deficit** are growing. "You cannot accuse Europe of being undemocratic. You might say it is an imperfect democracy, but with the crisis one has to recognise that there is an extraordinary sensation of **democratic deficit**," said a European source. (*Guardian* 25/04/2013)

Note the anonymity of attribution to a *European source*. On another occasion, it is blamed – by an MEP – on *Westminster* rather than the EU itself:

(8) We [in the European Parliament] do our best to hold the council of ministers to account, but a major part of the **democratic deficit** can be found in the way that national parliaments hold their ministers to account for their actions in Brussels. In Westminster we allow governments to get away too often with the cliched arguments of "we won the good stuff, but Brussels is imposing the bits we don't like" (*Guardian* 22/05/2013)

but on two occasions, the voice of the newspaper itself attests *the EU's democratic deficit* as reality, for instance:

(9) This means that a two-speed eurozone, divided between northern and southern states, becomes more likely, with Brussels and Berlin incurring rising unpopularity in the latter as anti-austerity election results underline the EU's **democratic deficit**. (*Guardian* 08/04/2013)

On the evidence of these key-items, then, the attitude of the UK's most influential pro-EU newspaper seems to have become more complex than we had anticipated at the outset of this research.

2.2 Qualitative analysis I: Concordancing metaphors and motifs

Tadros (1993) and Sinclair (2004) have both underlined the importance of distinguishing between *averring* (that is, attesting) and *attributing* opinions and attitudes, a distinction which is never more important than when studying political/media language. Moreover, it is clearly important to know whether an attribution is made in order to endorse or to dis-endorse and, as Sinclair puts it, to "challenge" the opinions being expressed. But these are distinctions which risk being blurred by the sort of statistical overview which corpus techniques employ. Only a qualitative analysis, close reading, can inform the analysts whether, for example, key items like *austerity* and *troika* are being used with approval or disapproval. In addition, only a close reading of *all* instances of mentions and uses of the notion will tell us proportionally how often the concept is averred and how often attributed and what the general evaluative stance towards the notion the newspaper adopts, that is, whether in general it endorses or challenges or is neutral towards it. In such cases qualitative and quantitative analyses feed into each other.

In this regard, it is instructive to investigate the *Mail*'s use of the template *claw* back powers* from the EU and the *Guardian*'s *repatriating powers* from Brussels, which are key-items in DM13_EU and GN13_EU respectively, and thus presumably a key concept in the newspapers discussions around the EU in 2013. However, a key-item list will not reveal whether the newspaper avers the concept or attributes it to other voices, nor what evaluative stance it adopts to the notion. The items were therefore concordanced (co-text 450 characters) for a closer reading in context.

The first thing to note is that, in the speech in which Mr Cameron announced the intention to renegotiate the UK's relationship with the EU, delivered on January 23rd 2013, he uses neither the expression *claw* back* nor *repatriat** powers. The term he actually used was *flow back*: "[…] power must be able to flow back to Member States, not just away from them" (Cameron 23/01/2013) but this wording is never found in either DM13_full or GN13_full. The choice of terms *claw back* and *repatriat** are therefore made independently by the newspapers themselves.

Taking the *Mail*'s *claw* back powers* template (22 occurrences in DM13_EU), it is noteworthy that the one other meaning-context in which the newspaper uses *claw back* with any regularity (16 occurrences in DM13_full) is in regard to ill-gotten bank profits and banker's bonuses. The item has a positive evaluative prosody (Louw 1993); it is right to *claw back* things which have been wrongfully taken away. The template is found nine times in GN13_full; five occurrences refer to clawing back bankers' bonuses and only four to powers from Brussels.

Concordancing of the template reveals that on one occasion the *Mail* endorses the idea quite explicitly in its own voice:

(10) The *Mail* has always supported the Prime Minister in his efforts to **claw back** powers from Brussels before holding a referendum […] (*Mail*, 11/05/2013)

and in the following headline, the positivity is still evident if more implicit:

(11) PM TO UNVEIL TORY BILL PROMISING A REFERENDUM – AND OBAMA PRAISES HIS PLAN TO **CLAW BACK** POWERS (*Mail*, 14/05/2013)

Other signs of positive endorsement of the notion are more subtle and embedded. Apart from the two cases above, the plan to claw back power is mentioned in apparently neutral simple news reports but the point of view taken is often that of the Prime Minister who *hopes to*, *wants to*, *promises to*, *is trying to*, is *determined to*, demonstrates *efforts to*, an *attempt to* claw back powers. Previous studies of the use of the items *effort/s to attempt/s to* in newspaper prose showed that they were generally efforts/attempts to achieve something praiseworthy (Sinclair 2004: 175–176, Partington, Duguid & Taylor 2013: 49). On the evidence of DM13_full and GN13_full, the same is true of *determined to* (*Mail*: *get even better value for money*; *Guardian*: *be a good father*, *be fair-minded*).

Turning to the *Guardian*'s *repatriat** of *power/s* template (53 occurrences in GN13_EU), the item *repatriate* is unlikely to have a positive evaluative prosody for its liberal-left readership, with its echoes of forced expulsion, an echo found in GN05_EU:

(12) But what to do about this new underclass – the "sans plastique", made up of foreign nationals who have no ID cards – could dominate the political agenda at the next election. Some, including the Conservatives, will be demanding their compulsory **repatriation**. (*Guardian*, 18/05/2005)

Concordancing the template in GN13_EU revealed that there is no entirely open and explicit endorsement or rejection of the notion of repatriating powers issued by the newspaper, although it does once claim that it "got a frosty reception across much of Europe". On four occasions we find the distancing technique of quotation marks:

(13) May announced her intention last October to withdraw from a series of policing and criminal justice measures under the banner of "**repatriating** British powers from Brussels" (*Guardian*, 23/04/2013)

and on another occasion the distancing *so-called*:

(14) Cameron argues that the EU is going to undertake a major treaty change anyway so no one will mind if, along the way, we ask for a recasting of Britain's membership, the so-called **repatriation** of powers. But he seems

> not to have noticed that Germany and other eurozone partners are wary of a major treaty revision. (*Guardian*, 07/01/2013)

Note the sarcastic appraisal of the Conservative leader in *seems not to have noticed*.

The *Guardian* rhetoric on *repatriating powers* is often more negative and antagonistic to the idea than the *Mail*'s on *clawing them back*, the former uses the term *demand/s* repatriation on thirteen occasions and *threat* on another four, for instance:

> (15) Cameron made his thinly veiled threat on the BBC's Andrew Marr Show, when he said he was "entitled" and "enabled" to seek a **repatriation** of powers […] (*Guardian*, 07/01/2013)

and the PM is even accused of *trying to put a gun to the head of his European partners*. This is of course, as much anti-Conservative rhetoric, quite naturally a *Guardian* staple, as pro-EU support.

On other occasions, however, the language is more neutral, for instance, *proposals* and *pledge* are used and space is even given to "pro-repatriation" voices:

> (16) JD Wetherspoon's chairman, Tim Martin, said the Europe-wide pub chain wanted "proper democracy" in EU countries, where accountable local politicians make laws. "The idea that we are in trouble if we **repatriate** democratic powers is piffle," Martin said. (*Guardian*, 18/01/2013)

In this regard, the fact that newspaper such as the *Guardian* often publishes voices whose views are at odds with the official editorial line, tells us that there is a third potential media stance towards a political issue besides averral and attribution (2.3.1), given that the latter is frequently a convenient mechanism of ascribing a media outlet's own beliefs to a third-party. The newspaper does not aver or promote some types of views but, by choosing to publish them, it acts as a *vehicle*. There are three options, then, in news media language stances: an opinion may be *averred*, *attributed* or *vehicled* (the OED has an entry for the word as a verb, but some might prefer *channelled*).

2.3 Qualitative analysis II: Leading articles

On the evidence so far, there appeared to be some doubt as to whether the *Guardian* in 2013 was still a pro-EU voice at all, given the number of conflicting opinions it carries. Had the trauma of the Eurozone crisis shaken its pro-EU stance to the point of abandonment? To answer this question we conducted a further qualitative analysis, namely, close-reading of the newspaper's own, open and official voice, its explicitly averred rather than attributed or vehicled opinions on the EU. We decided to read the *Guardian*'s editorials to be found in GN13_EU, given that:

> Editorials are the voice of the newspaper […] One of the prime functions of editorial comment is that of persuading the newspaper's readers of its point of view [and] The editorial tries to persuade first of all through its authoritarian stance. (Morley 2004: 239)

The *Guardian* calls its editorials 'leading articles' and GN13_EU contained 21 (13,400 words); thus these were all official statements issued by the newspaper dedicated at least in part to the EU and normally dealing with the UK's relationship with it.

Here finally was evidence of the paper's pro-EU stance, although this evolves somewhat over the nine-month period. The repeated message is that Britain must "engage", play a more central role, enter metaphorically into the *engine room*:

(17) In this new German-dominated Europe, where is Britain? Britain may never have placed itself in the very engine room of the European Union, as France chose to do. But does Britain have to continue to sulk in a dinghy being towed along by the main vessel? (*Guardian*, 03/04/2013)

and the role of all the UK's political parties should be to work towards *a more perfect union* (*Guardian* 31/05/2013, with an obvious reference to Senator Obama's famous 2008 election speech). There is also sarcasm at the beginning of the year about *Cameron's hokey-cokey*: *In-out EU referendum* (headline, *Guardian*, 24-01-2013).

The editorial voice blames EU-scepticism on xenophobia, *politicians* and on *rightwing newspapers*, *many of whose owners do not pay tax in this country*, perhaps a swipe at Rupert Murdoch's News Corporation:

(18) For a variety of reasons, in which history, geography, culture and language are intertwined, and which include remnants of a postcolonial self-delusion about British superiority and continental inferiority, many British people are reluctantly and half-heartedly engaged with Europe. Partly for that reason, too many politicians of all parties find it easier to parrot or appease the views of a few rightwing newspapers, many of whose owners do not pay taxes in this country and regard "Europe" as synonymous with regulations which threaten their interests as owners and rich people. (*Guardian*, 11/01/2013)

It perhaps needs to be said, however, that the *Guardian* itself has not paid corporation tax in the UK for a number of years either, because it runs at a considerable loss.[5] Wisely, the *Guardian* refrains from blaming the *public* for EU-scepticism, which it portrays as savvy but let down by authority:

5. <http://www.gmgplc.co.uk/gmg/faqs/#q9> (16 August 2018)

(19) Many members of the public are instinctively more cautious and more pragmatic, not least because they do not trust the press, but they get little lead from politicians. (*Guardian*, 11/01/2013)

and note yet again the theme of "we, the beleaguered voice":

(20) The result, over many years, has been the growth of an often banal anti-European populism on the right and in parts of the left. This has now generated something of a Eurosceptic hegemony in British public debate. (*Guardian*, 11/01/2013)

and of the dangers *[t]his country* faces if our voices do not speak out:

(21) This country is at risk of allowing itself to be stampeded by the Tory party and the Europhobic press into abandoning its place in Europe. Pro-Europeans should shed their anxieties. Voices that have been silent for too long need to make themselves heard. (*Guardian*, 10/01/2013)

In general, at the start of the year at least, who is *not* to blame for EU-scepticism is the EU itself, but as the year progresses, even the editorial voice of the *Guardian* begins to have a problem, that of supposedly EU-imposed *austerity* (mentioned 20 times in the leading articles and, as we saw above, a GN13_EU keyword). The left-leaning *Guardian* is opposed to austerity and the imposition of tough sacrifices on vulnerable sections of society, and recognises how damaging the policy is for pro-EU voices: "across the continent, the EU's authority is taking a battering, as austerity becomes transnational" (*Guardian* 31/05/2013). Nor does such seemingly authoritarian imposition sit well with the idealistic vision of harmony, cooperation and general niceness of the EU often held by its supporters:

(22) Now these same, broken economies are being fed not Frankfurt's credit but austerity Berlin-style – and they cannot cope. (*Guardian*, 27/03/2013)

And so even the *Guardian*'s pro-EU leading article voice is conflicted. The EU was not supposed to generate such asymmetry of power:

(23) […] the European Union, acting overwhelmingly at the bidding of Germany, has imposed humbling terms on Cyprus. In that light – and in the light of similar crises in Italy, Spain and Greece – it could be foolish to deny that a new form of single-force domination is emerging […] The Financial Times columnist Gideon Rachman put it succinctly yesterday: "Growing German power – and growing resentment of that power – are now the main themes in European politics […]" (*Guardian*, 27/03/2013)

Elsewhere in the *Guardian*, austerity is described as *German-led*, even though the so-called 'Troika' of lenders supposedly imposing austerity is not specifically

German, being composed of the IMF, the European Central Bank, and the European Commission. The concordance of the *troika* reveals that the anti-austerity 'Fuck the Troika' movement receives ample and sympathetic treatment and also one of the most damning indictments of the current state of the EU in either newspaper:

(24) […] the modern EU has [given] a failed neoliberal model of capitalism the force of treaty, entrenching deregulation and privatisation and enforcing corporate power over employment rights. Claims that the single market would boost growth have proved groundless. But the EU's profoundly undemocratic and dysfunctional structures have been brutally exposed by the eurozone crisis and the devastation wreaked by **Troika**-imposed austerity. (*Guardian*, 15/05/2013)

accompanied by support for the Conservative plan for the referendum which elsewhere in the paper, as we saw, was mocked as Cameron's *hokey-cokey*:

(25) The fallout from that crisis means the EU will in any case have to be restructured. Given those circumstances and the Tory commitment, it would be both wrong in principle and politically foolish for Labour not to back a referendum. (*Guardian*, 15/05/2013)

The writer in (24) and (25) is Seumas Milne, probably the *Guardian*'s most left-wing columnist, who, in 2015 became Director of Strategy and Communications for the EU-sceptic Labour Party leader, Jeremy Corbyn. A reminder that anti-EU sentiment in the UK is, and never was, by any means confined to the right.

2.4 Who are we up against?

In his 2001 piece, Teubert laments the lack of a bold and authentic pro-EU voice in the UK media, whilst we have seen repeated claims from the *Guardian* of being a lonely pro-EU voice struggling to be heard against a howling British "Europhobic" press.

Unsurprisingly, this is not how the rightwing press sees the world. They see the might of national and international institutions, including all the main political parties and the media, ranged against them. Marchi and Taylor (2009) noted the following passage from the right-leaning *Daily Telegraph* editorial extract below, published after a referendum rejecting a European Constitution in 2005, under the headline *Mere democracy won't stop the EU machine*:

(26) In defiance of a united media, a monolithically pro-Brussels political class and blizzards of propaganda, they have said a resounding *Non* to the *Euroelites* who have governed them for half a century. (*Telegraph*, 30/05/2005)

whilst a *Daily Mail* headline in 2013 underlines a supposedly vast EU budget for self-promotion, sarcastically claiming:

(27) EU's £2.4bn ad budget is bigger than Coca-Cola's (*Mail*, 30/09/2013)

and special spleen is vented on that nest of supposed pro-EU vipers, the BBC:

(28) LABOUR'S former Euro Trade Commissioner Lord Mandelson was invited by the BBC to defend the EU for Radio 4 yesterday – just as David Cameron was giving a speech about renegotiating our membership. An indication of the **BBC**'s pro-EU tendency? (*Mail*, 24/01/2013)

(29) The **BBC** may be unable to see it, but the British economy appears at last to be stirring into action. And, much as the Corporation yearns for it to be otherwise, Euroland is not. (*Mail*, 15/08/2013)

2.5 Representing *Eurosceptics*, *Europhobes* and other *Euro*-animals

In this section, we look specifically at how the two newspapers talk about *Euroscepticism*, *Eurosceptics* and similar items.

2.5.1 *The* Guardian

The *Guardian*'s attitude to *Euroscepticism* is, predictably, largely negative but also more nuanced than expected. A concordance of *eurosceptic** in GN13_full produced 217 occurrences. There is often open derision: *Eurosceptic Gollums*, *noisily Eurosceptic*, *gaffe-prone Eurosceptic populists*, *rabidly Eurosceptic renegades*, *poster boy of the Tory Eurosceptics*, *the provisional wing of the Conservative party* (quoting Labour's Mr Mandelson, who likens sceptics to violent IRA militants). At other times the negativity is more veiled: banging the *Eurosceptic drum*, *the Eurosceptic agenda*, *rising tide/wave of Euroscepticism*.

However, there is also the occasional recognition that not all Eurosceptics are tarred with the same brush:

(30) The Conservative party's divisions on Europe lie between those **Eurosceptics** who want to leave the European Union and those **Eurosceptics** who, like David Cameron, would like to negotiate a way of remaining part of it. (*Guardian*, 30/09/2013)

The concordance of *europhob** in GN13_full produced 15 occurrences; Europhobes are referred to derisively as *militant*, *lunatic*, *attack-labrador*. Both concordances contain once again signs of a conviction that our side's voice is a lonely, beleagured one:

(31) This has now generated something of a **Eurosceptic** hegemony in British public debate. (*Guardian*, 11/01/2013)

(32) The speech's real concern, however, was not economics but politics – the politics of a restive Tory backbench, an insurgent Ukip and a mostly **Europhobic** press. (*Guardian*, 24/01/2013)

2.5.2 *The* Mail

What comes as some surprise, after concordancing these items in 'DM13_full', is that the *Mail*'s attitude to *Euroscepticism* and *Eurospectics* is by no means always positive, in fact the nearest they receive to open praise is in descriptions such as *veteran*, *heavyweight* and *ancient seer* of *Euroscepticism*. The following call to arms is the only example of open encouragement:

(33) The aim [of a pro-EU campaign] is to scare the living daylights out of the majority of Britons who are fed up with the corrupt, repressive hegemony of Brussels. All in the hope that the UK will vote to keep the eurocrats' gravy train rumbling on. In the interests of a balanced debate, **eurosceptics** must get organised without delay. For they are up against some of the best-connected and most unscrupulous spinners in the business. (*Mail*, 25/07/2013)

Note yet again how "our voice", this time the *Eurosceptic* voice, is portrayed as underdog. But the *Mail* also describes Eurosceptics as *strident* and they are often portrayed as irascible, quick to *fury* and *anger*, easily *infuriated*, *enraged*. There is even the following highly sarcastic reference:

(34) PREPARING for a showdown with his party's Europhiles, the Conservatives' **Eurosceptic** bore-in-chief Bill Cash quotes Cassius from Shakespeare's Julius Caesar, declaring defiantly: We'll meet them at Philippi! TV classicist Professor Mary Beard asks drily: Does he realise Cassius lost the battle of Philippi? (*Mail*, 16/01/2013)

The complexity posed by Eurosceptics for the Tory-leaning *Mail* is that the anti-EU bloc inside the Conservative party are *rebellious* and *mutinous* and threaten the party's stability, whilst the openly anti-EU UKIP "who are far beyond scepticism" (*Mail*, 14/01/2013) threaten its electoral success by attracting EU-sceptic votes.

The term *europhob** is found just once in DM13_EU and it is not in the newspaper's own voice; it is attributed and balanced by reference to *Europhiles*, seen as matching if opposed extremes:

(35) '[…] Businesses reject both the Europhile dream of further integration and the Europhobe dream of a complete exit from the EU, provided a satisfactory renegotiation is achieved,' said Mr Longworth.

2.6 Metaphors and evaluation

A final observation on the 2013 EU discourses is how rich a source of metaphors they are, as indeed was noted by Mr Cameron in his January speech:

(36) Let's stop all this talk of two-speed Europe, of fast lanes and slow lanes, of countries missing trains and buses, and consign the whole weary caravan of **metaphors** to a permanent siding. (Cameron 23/01/2013)

although to use a metaphor in complaining about metaphor is certainly rhetorically interesting.

Metaphors in discourse are almost always used to evaluate and in political and media discourses they are part of the rhetoric of praise and blame (Partington & Taylor 2018). A *pioneer Europe* (*Guardian*) is a good thing as is being in the *engine room* and at the *top table* of the EU (both *Guardian*), whilst the *Brussels yoke* (*Mail*) and *a cosmopolitan transcontinental ménage* (*Guardian*) are both bad things. *Clawing back* powers is praiseworthy, whilst the UK is very much to blame for *sulking in a dinghy* behind the main EU *ship* (*Guardian*). The *Guardian*'s metaphorical photo caption: *Angela Merkel, the prudent housewife* (01/04/2013) is meant to be mocking, given the paper's aversion to *Teutonic austerity*, but succeeds only in being sexist and demeaning; an alternative positive evaluation might have been *Dr Angela Merkel, PhD (quantum chemistry)*. There is a body of literature on metaphors of the EU (for example, as a *house* or *building*, planned by *architects* and with solid or shaky *foundations*), many of which are discussed and referenced in Musolff (2017).

In their investigation of UK newspaper discourses around the euro, Vaghi and Venuti (2004) noted how the Eurozone and EU were often described using terms like *in*, *inside*, *outside*, *within*, denoting what Lakoff and Johnson (1980: 29–32) call the CONTAINER metaphor, a conceptual or existential construct. Duguid (personal communication), instead, reads many of the terms – *entry*, *joining*, *membership* – as denoting a hyper-metaphor of the kind the EURO IS A CLUB, a cultural rather than existential construct. Indeed, the UK Socialist Labour Party's website calls the EU *a capitalist club*.[6]

We decided then to concordance in DM13_EU the items *leav** and *exit* in a 5-word span of *EU* and *European Union*, both of which could pertain to either the

6. <http://www.socialist-labour-party.org.uk/europeanunion.html> (16 August 2018)

CONTAINER or CLUB hyper-metaphors. There were 52 occurrences, all of which were attributed to voices engaged in discussions about the benefits and dangers of exit and co-occurring vocabulary includes reporting words, such as *said*, *tell**, *insist**. There is not a single occurrence of the *Daily Mail*'s own voice averring an opinion on whether or not to leave the EU. Although keen to support the in-out referendum, the paper's authorial voice, perhaps in the spirit of being careful what you wish for, is not prepared to commit itself, in contrast with the *Guardian*'s editorial voice, which, as we saw, was in favour of membership and even further integration.

Very surprisingly, there were twice as many negative as positive voices vehicled by the *Daily Mail* (27 to 14), that is, voices *warning* (11 occurrences, the most common reporting item) that exiting would be *mad*, *reckless*, a *historical error*, *catastrophic for trade* and that it would leave Britain *isolated*, *weakened*, condemned to *irrelevance on the world stage*. Along with *world stage* we find other metaphors in play: not *sliding* or *sleepwalking* (a word with negative evaluative prosody) towards the *exit door*. The item *club* is used explicitly, in a *warning* by the President of the European Council, Barroso, that Britain would pay a high price for *leaving the club altogether*, presumably with a very different evaluation of the EU IS AN (EXCLUSIVE) CLUB metaphor than that of the Socialist Labour Party noted above. The pro-EU voices vehicled by the *Mail* rely heavily on a rhetoric of fear (see Section 3.1.2 below) of the consequences of change, as of course is normal in any political campaign in favour of a status quo (Brader 2006).

2.7 Conclusions on the 2013 discourses

The main conclusion from this 2013 study is that it is unrealistic to dichotomise sentiments about the EU straightforwardly into pro-EU and anti-EU voices, even though sometimes media voices themselves do so. In this regard, we need to be careful of the dichotomising tendency of some comparative methods of research, especially keywords procedures, which just highlight differences (Rayson 2012, Taylor 2013); it was in the more qualitative analyses, for instance, in the concordances of *eurosceptic**, that the considerable overlap was found between the two newspapers' stances.

There are occasional full-blooded anti-EU voices on the right and occasionally on the left but not *all* criticism should be seen as anti-EU. As the *Guardian* itself points out, "Euroscepticism is a broad church" and, indeed, the *Guardian* itself occasionally worships there. In general, then, there seem to be at least four main viewpoints:

1. the die-hard pro-EU idealists, defensive of the EU, offended and puzzled by criticism of it and at most admitting "the country is undoubtedly cross about Europe just as it is cross about Westminster and much else besides" (*Guardian* leader, 24/01/2013).
2. the EU optimists, conscious of some need for reform but for whom the solution is closer engagement and *more Europe* (a *Guardian* key 2-gram). For these first two groups the real problem is the UK, and if only the UK took more responsibility (entered the *engine room*) it would be the *solution* to the EU's woes.
3. the EU "looseners", keen to renegotiate institutions and treaties "not fit for purpose" (*Guardian* 16/01/2013), generally wanting the UK to engage with the EU but on an ad hoc, issue-by-issue basis.
4. the anti-EU voices who see no good at all in the EU and cannot get out of the door fast enough.

The first and last groups are, naturally, the most aggressive. Whilst the favourite butts of the latter are the *Eurocrats* of *Euroland*, inept and corrupt in equal measure, the targets of the former are a caricaturised set of EU-phobic fanatics.

Two final reflections are offered on the negativity of much of the EU discourse in these two newspapers in this period. Firstly, negativity is, of course, one of the most prominent news values (Galtung & Ruge 1965) and is salient in political reporting. The actions of powerful institutions such as Westminster or the US administration come under – many would argue *must* come under – a great deal of scrutiny and criticism. There is no reason why the institutions and administrators of the EU, hardly a disempowered minority, should be exempt.

Secondly, discourses, especially the composite media discourses we have analysed here, do not generally arise and develop in a vacuum or uninfluenced by real-world events, and EU-sceptic discourse is again no exception. Some EU-scepticism may well be a product of antagonistic reporting, but had monetary union been an unqualified success, had the Eurozone not entered into a long-lasting crisis in 2008, and had the EU response not been to adopt austerity measures, we might reasonably have expected the discourses around the EU, in the *Guardian* at least, to be more positive. Baker (2012) introduces and discusses the notion of what constitutes "acceptable bias" in news media reporting. We would prefer the term "acceptable" or "reasonable *negativity*", since "bias" strongly implies prejudice and partiality. But this notion raises the questions of, first, acceptable or reasonable to *whom* – surely not just to an individual analyst – and, second, what are the real-world events and processes which might be causing negative reporting. As Baker maintains, these are questions which corpus linguistics techniques can usefully address, especially given its inbuilt capacity to compare different corpora and therefore discourses on different entities (and, as here, discourses on the same

topic from different sources). As Gries (2011) argues, it is not methodologically or even deontologically valid to make conclusions about media negative representation of – his example – one religious group without comparing how the same media sources represent other religious groups (Gries 2011: 97). Similarly, it would not be justifiable for us to make judgements on "(un)reasonable bias" towards a political entity like the EU unless we had taken the representation of other political bodies into account.

Finally, here, if we compare the fragments of EU-sceptic discourse analysed here with those in Teubert (2001), we note first of all the notion that sovereignty is not in danger of being removed but *has* been removed and must be taken back. We see too the disappearance of preoccupations over loss of *Anglo-Saxon* or *transatlantic* values, or the *Commonwealth*, which are barely mentioned even in DM13_EU. But new worries have arisen, fears of mass immigration from new member states, of a German political and economic hegemony and of EU-imposed austerity politics.

3. 2016: The campaigns immediately before the vote and the reactions just after

This third study examines how attitudes to the 2016 Brexit Referendum were represented in a number of UK national newspapers in the three-month period immediately before the vote (22 March 2016 – 22 June 2016, the 'pre-Brexit corpus') and in a similar stretch of time after (24 June 2016 – 24 September 2016, the 'post-Brexit corpus'). All articles containing in their headline the words *Brexit* OR *EU Referendum* were collected from four daily British newspapers: the *Daily Telegraph*, the *Daily Mail* (with its Sunday equivalent the *Mail on Sunday*), the *Guardian* and the *Daily Mirror* (with the *Sunday Mirror*). We chose two right-leaning and largely pro-Leave newspapers (a broadsheet and a tabloid) and two left-leaning and largely pro-Remain ones (again, a broadsheet and a tabloid) in order to encompass a cross-section of UK news reporting, with newspapers of opposing political stances and of tabloid and broadsheet reporting styles. Using these parameters, 3,290 articles were included in the corpus amounting to about 2.5 million words (Table 4). The corpus was examined using *WordSmith Tools* suite (Scott 2008).

Table 4. Composition of the Brexit News Corpus

Newspaper titles	Number of articles		Number of words (approximate)	
	Pre-Brexit	Post-Brexit	Pre-Brexit	Post-Brexit
The *Daily Telegraph*	281	492	169,000	316,500
The *Guardian*	896	1003	810,000	866,500
The *Daily Mail*	109	159	74,000	103,000
The *Daily Mirror*	116	234	116,000	87,000
TOTAL	1402	1888	1,169,000	1,373,000

3.1 Findings

In this section we firstly discuss the media coverage in the lead-up to the referendum. In the pre-Brexit corpus, three main themes were indicated by the frequency word lists. The first is concerned with the pre-vote representation of the referendum itself. The second is the widespread so-called scaremongering of which both camps were accused and the third theme was immigration.

3.1.1 *The representation of the referendum*

Concordancing the item *referendum* in the pre-Brexit corpus revealed 3,105 occurrences. Of these, it was part of the construction *pre-referendum* in 22 cases, of *post-referendum* in 21 and of *referendum-related* in 14.

We made note of the adjectives that were used to describe the event. Some of them did not betray the political stance of the writer – or of the person quoted – but only acknowledged that the referendum was happening soon (*forthcoming*) or that it was objectively an event of crucial importance (*historic*, *landmark*). However, the majority of the adjectives expressed quite strong evaluation.

Among the adjectives with a positive connotation is *long-delayed*, used by Boris Johnson, the former Conservative mayor of London:

(37) So when the British had their **long-delayed** referendum, in June 2016, they were being offered the worst of both worlds. (*Telegraph* 23/05/16)

similar to *long-delayed* is *long-awaited*:

(38) […] and back in Britain for the **long-awaited** referendum on June 23. (*Telegraph* 28/05/16)

Surprisingly, these were the only two adjectives in the corpus expressing a totally positive evaluation of the referendum. All the others had a more or less explicit negative connotation.

The pro-Remain *Guardian* in particular used a broad range of negative adjectives: from the more moderate critical and divisive to the highly accusatory intemperate, *reckless*, *rancorous*, *accursed*, *wretched* and even *damn*. Here some of the most explanatory sentences:

(39) Cameron is to blame for unleashing this **reckless** referendum, but above all for the anger seething underneath it. (*Guardian* 20/06/16)

(40) There was football on the other side, bottles of wine beckoning from fridges, and this **damn** referendum already feels like it's gone on for ever. (*Guardian* 03/06/16)

The use of the adjective *looming* (30 occurrences) is interesting:

(41) Fears over the **looming** European Union referendum may have put the brakes on the UK jobs recovery, official figures are expected to show this week. (*Mail on Sunday* 17/04/16)

(42) The Bank of England has not seen any signs that UK companies are scaling back their investment plans because of uncertainty caused by the **looming** EU referendum, a top Bank official said. (*Guardian* 05/05/16)

While the OED simply defines the verb *to loom* as "coming indistinctively into view", further examples from newspaper articles suggest that looming is often something threatening:

(43) But the barriers to tourism are daunting: dozens of heavily-armed militias, a desperately weak central government, jihadi terrorism and, some warn, the **looming** threat of state failure. (*Guardian* 7/11/13)

In newspaper prose at least, looming has a negative evaluative prosody and its use in association with the referendum betrays the writer's negative attitude.

3.1.2 *The theme of fear*

A considerable number of items in the word-frequency lists pertained to the emotion of fear, and so a concordance was prepared from the whole pre-Brexit corpus of the search term *fear**, which yielded 676 occurrences, including the items *fear*, *feared*, *fearful*, *fearing*, *fearless*, *fears*, *fearsome* as well as *fear-mongering* and *fearmongering*.

The most frequently co-occurring items were, unsurprisingly, grammatical word such as *of*, *to*, *that*, *about*. Beyond this, other words could be attributed to one of the following categories:

1. the referendum (*Brexit, vote, leaving*, …)
2. the campaigns (*Leave, Remain, Project Fear, tactics*, …)

3. the economy (*Business, economic slowdown/growth, market, investments, sterling*, …)
4. immigration (*immigration, fear of the other*, …)
5. the people involved (*voters, British, government, Osborne*, …)
6. areas involved (*UK, Britain, Ireland, global*, …)

What emerges from this first categorisation of the terms most often found in close proximity to discourses of fear, is that fear characterised both sides. The Remain side mainly tried to scare voters by emphasising the economic damage that Brexit would cause. The Leave side played on the fear of immigration. The resort to accusations of fearmongering or, more commonly, scaremongering on both sides was well documented in the press coverage. As a next step, then, we concordanced *scaremonger** and *scare-monger** in each newspaper subcorpus: the occurrences are reported in Table 5.

Table 5. Occurrences of *scaremonger*/scare-monger**

	scaremonger*	***scare-monger****
The *Mirror*	10	2
The *Daily Mail*	21	1
The *Guardian*	64	–
The *Telegraph*	35	–

Scanning through the lines individually, it became clear that *scaremonger* was rarely used as a verb, or as a noun to describe a person who deliberately spreads alarming reports. The majority of the concordance lines contained the form *scaremongering*, both as a noun indicating the practice of spreading alarming reports – and as an adjective, collocating with *campaign* and *tactics*.

The occurrence of pre-modification indicated various types of scaremongering. While the Brexit camp's scaremongering was once defined as *desperate* by the opponents, the adjectives describing the actions of the Remain camp gave a more complete idea of how their arguments were perceived. The Remainers' scaremongering is constructed as continuous (with collocates like *constant* and *incessant*), but also as unreasonable, bordering on crazed (*baseless*, *demented*, *ludicrous*). In addition, it is perceived as deliberately shocking (*blatant*, *cynical*, *naked*, *outrageous*), implying that some Remainers at least were telling the British people downright lies.

Looking at the agency and the actors, who accuses whom of scaremongering? The majority of the accusations in the pro-Leave press were directed towards the Prime Minister David Cameron, prominent ministers such as George Osborne

or the Chancellor Philip Hammond, the Government generally, the metonymic *Downing Street* or, more comprehensively, simply the *Establishment*.

They are mainly accused of spreading alarmist stories on Britain's economy after Brexit. Even when the Remain camp's claims are backed by experts, they are nevertheless dismissed as false and scaremongering, in the following case with a touch of sarcasm:

(44) GEORGE Osborne was accused of 'outrageous' **scaremongering** last night after suggesting pensioners would lose up to £32,000 each if Britain votes for Brexit. […] The Prime Minister and Chancellor have, in recent months, claimed Brexit could lead to war, genocide, recession, migrant camps in Kent, 800,000 job losses, house price collapse, stratospheric rises in clothing and food prices and the end of cheap holidays (*Daily Mail* 27/05/16)

Unsurprisingly, in the pro-Remain press there were many instances of the term being used against the Leave supporters. In the majority of these cases, the allegation is that of spreading false myths about immigration (see below):

(45) The threats coming from the Brexiteers about Turkey joining the EU and millions of Turks coming here are **scaremongering** in the extreme (*Daily Mirror* 7/06/16)

We then also searched for **monger** to examine other possible uses of the suffix *-monger*/*-mongering* in addition to *fearmonger* and *scaremonger*. The other term found was *doom-monger*/*doom-mongering*. The adjective *doom-mongering* always pre-modified predictions being made – according to Brexiters – by Remainers exaggerating the risks of Brexit:

(46) In recent days a string of Cabinet ministers have been wheeled out to make doom-**mongering** predictions about the risks of Brexit. (*Daily Mail*, 29/03/16)

(47) The warning is the latest in a series of **doom-mongering** predictions from pro-EU ministers. (*Daily Mail*, 27/05/16)

Overall, it seems clear that both camps used fear to make their case, but, interestingly, only the Remain campaign was successfully branded 'Project Fear', which could imply that the Brexiters' rhetorical campaign strategy was more effective.

3.1.3 *Discourses on immigration*

Finally, we looked at how the theme of immigration was dealt with in the pro-Leave and pro-Remain press by examining the articles in the pre-vote corpus containing the search word *immigra**, which yielded 1,058 occurrences. It was very often explicitly framed as one of the key themes of the EU referendum debate, especially

with reference to the Leave campaign. The majority of the discourses focused on the need to control immigration since (mass) immigration is a problem, something that was causing worries, anxieties and fears to the British people.

However, many articles noticed that the pros and cons of immigration were difficult to spot amid the confusing mix of promises, lies, nonsense fuelling the debate. The Government was even accused of "hiding the truth about immigration" (*Daily Mirror* 03/04/16).

The concordance lines and more detailed study of the cotext revealed a wide range of different types of representation. For example, there is a distinction between different types of immigrants based on their country of origin. At the most general level, immigrants are divided into EU or non-EU, but more specific distinctions are also present:

(48) But pulling up the drawbridge against the rest of Europe is the wrong answer. The right answer is the same one we used when migrants from **Ireland** were vilified in the last century; when **Jewish** immigrants were targeted a century ago; and when **Asian** and **African-Caribbean** workers were attacked in the 1950s, 60s and beyond. That is, strong trade unions delivering the rate for the job, whoever you are and wherever you come from. (*Guardian* 21/06/16)

(49) Columnist Trevor Kavanagh took up the immigration point: "The prospect that scares the pants off voters is mass **Muslim** immigration, now running at well over a million each year into Europe from **Pakistan**, **Afghanistan**, **Africa** and the **Arab** world." (*Guardian* 18/04/16)

Furthermore, immigrants are divided into categories also depending on their working skills – *professional*, *skilled*, *un-skilled* or *jobless*. *Illegal immigra** is mentioned on 23 occasions.

However, what also emerges as a widespread theme, especially in the Remain press, is *anti-immigration* (25 occurrences). The adjectives *racist*, *inflammatory* and *aggressive* are used to describe the kind of rhetoric allegedly being used by Brexiters. Vote Leave representatives – Farage, Gove, Johnson – are accused of spreading anti-immigrant *sentiment*, as is the Prime Minister, Cameron by the *Guardian*:

(50) For politicians to promise policies they cannot deliver disastrously undermines faith in the democratic system. That lack of trust is fuelling the leave campaign. But Cameron's pledge also fuelled **anti-immigration sentiment**. He framed mass immigration as a huge problem that required a radical solution: a target that would have achieved a monumental reduction in the number of people entering the country. (*Guardian* 21/06/16)

To cross-check and avoid being accused of cherry-picking, we looked for any evidence of attempts to talk about immigration from a more positive perspective. The *Guardian* reports opinions in favour of immigration, for instance, Hilary Benn, the Labour shadow foreign secretary made "a passionate case in favour of the benefits of immigration" (*Guardian* 13/06/16) and Andy Burnham, the Labour shadow home secretary, "gave a firm defence of the role immigrants play in the NHS" (*Guardian* 20/06/16). The *Daily Mirror* reports data from Migration Watch:

(51) A report by Right-wing pressure group Migration Watch will admit today the financial benefits of recent **migration** from the original 14 EU countries outweighs any cost incurred from the numbers arriving from Eastern Europe. […] A second report by the academics cited by **Migration Watch** also said EU migrants have made a positive contribution to the UK economy. (*Daily Mirror* 20/06/16)

3.2 The result and post-vote Britain

In this section we first discuss the first reactions in the media coverage followed by separate discussions of pro-Leave and pro-Remain reactions.

3.2.1 *First reactions*

At this point, an analysis of the post-Brexit vote corpus was necessary to understand how the Leave victory was received by the press. How was the event reported immediately after the vote had been cast? In the post-Brexit corpus, we looked for collocations of *result*:

(52) On the pro-Leave side, the *Daily Mail* wrote 'The historic **result** could see us embarking on a path to an enlightened era of prosperous global trade, freed from the shackles of unelected Brussels bureaucracy' (*Daily Mail* 24/06/16)

We noted an array of positive evaluations, *historic*, *enlightened*, *prosperous*, *freed*, and even *embarking* which has a positive evaluative prosody. For the Leave side, then, after getting over their initial surprise at having actually won, the result brought optimism about Britain's future ahead of Brexit.

Furthermore, the *Telegraph* emphasised its belief that the economic consequences of the vote were not as brutal as experts had predicted, indeed, quite the opposite:

(53) The referendum result caused pandemonium in the markets – and gave rise to opportunities. (*Daily Telegraph* 02/07/16)

(54) Markit said the fall in the value of the pound triggered by the referendum result had helped to push up overseas orders, while domestic output also bounced back and employment rose for the first time this year.
(*Daily Telegraph* 02/09/16)

On the losing Remain side, the *Guardian*, in a piece published in the early hours of June 24, when the results of the vote were clear, referred to Brexit as an *earthquake*, t*he rubble [from which] will take years to clear*. The majority of the articles expressed a sense of uncertainty and disappointment generated by the result. In some cases, the outcome was defined as *bruising*, *cataclysmic*, *disastrous* as well as *a shock*.

The next step was to identify the most discussed themes in the pro-Leave and pro-Remain newspapers over the full three months following the vote; to this end the *WordSmith* Keyword tool was used to compare the key items of the two sub-corpora.

3.2.2 *Pro-Leave reactions*

Table 6, lists a selection of fifteen terms that were significantly more frequent in the pro-Leave newspapers than in the pro-Remain, all appearing among the top 120 items.

Table 6. Fifteen selected pro-Leave keywords

business, companies, customers, investors, market, trade, doom, deals, boost, shares, sales, opportunities, income, growth, invest

The majority of the terms are economy-related. This might come as a surprise given how the destiny of the economy was alleged, by the Remain campaign, to be the Achilles' heel of the Leave campaign. We then concordanced in greater detail two words of these items with a strong positive connotation, *boost* and *opportunities*.

The word *boost* co-occurs with items relating to the economy, *growth*, *industry*, *investment* or *sales*. A *boost* in British tourism is also mentioned. Furthermore, the term is often found in the collocation *Brexit boost* or *post-Brexit boost*, emphasising the Leavers' conviction that the boost was a direct consequence of the vote:

(55) **Post-Brexit boost** for UK as China hints at trade deal (*Telegraph* 07/07/16)

(56) The staggering extent of the **post-Brexit boost** to tourism in Britain is revealed today in official figures showing billions more pounds flowing into the industry. (*Mail on Sunday* 21/08/16)

The concordance of *opportunities* confirms the Brexiters' belief that the June vote opened new possibilities of improvement for Britain in general and especially for its economy, in sharp contrast to the pre-referendum dark warnings of the Remain

campaign. The adjectives describing these opportunities are extremely positive: *desirable*, *fantastic*, *huge*, *significant*, *tremendous*:

(57) Leaving the EU presents **tremendous opportunities** to develop new agricultural, fisheries and environment policies tailored specifically to the industry and landscape of the United Kingdom, as well as our wildlife […] (*Telegraph* 16/09/16)

Britain and the British people are encouraged to seize these opportunities and make the most of them.

Among the pro-Leave keywords was also the negative term *doom*. It is used to emphasise the fact that the negative predictions made by Remain supporters of the consequences of Brexit were supposedly not fulfilled after the vote:

(58) The Bank and the Treasury have lost credibility; By trumpeting unfulfilled prophecies of **doom** about Brexit, they have reduced our faith in their judgment (*Telegraph* 09/09/16)

In a smaller number of cases, doom appeared in the constructions *doom-mongering* and *doom-laden* and in the expression *doom and gloom*, always with reference to the pre-vote *warnings* of Remainers:

(59) The forecasts represent a climbdown for the global financial watchdog [the IMF] after it issued a string of **doom-laden warnings** over the damage Brexit would do. (*Daily Mail* 20/07/16)

The economy, then, constitutes the core of the discussion in the pro-Leave newspapers. To a lesser extent, there is also an interest in undermining the credibility of those authorities which imagined a doomsday scenario that has in their view not come to pass.

3.2.3 *Pro-Remain reactions*

In contrast to the pro-Leave newspapers, in the pro-Remain press what emerges clearly is a focus on the putative social consequences of Brexit rather than on the economic ones. Table 7 shows a selection of keywords in the pro-Remain corpus, all appearing among the top 120 items. These words suggest that attention was paid to the alleged episodes of racism which occurred in the immediate aftermath of the vote.

Table 7. Fifteen selected pro-Remain keywords

feel, social, xenophobic, Polish, racist, community, racism, xenophobia, border, inequality, Muslim, attack, fight, incidents, feelings

A closer analysis of the concordance lines of *racis** and *xenophobi** in the pro-Remain corpus showed that the two terms often co-occurred, showing how the two concepts are perceived as connected. The adjectives *racist* and *xenophobic* very frequently collocated with words like *abuse*, *attack*, *crimes*, *disorder* or *incidents*. The referendum result – together with the pro-Leave campaign rhetoric of the previous months – is thus reported as an event that fuelled a climate of racism and xenophobia.

In the key word list, the terms *Polish* and *Muslim* suggested who the targets of these attacks were said to be, as confirmed in the fragment below:

(60) The National Police Chiefs' Council said harsh sentences would be handed down to anyone convicted of racist disorder. It comes after **Polish** and **Muslim** communities reported being targeted. (*Daily Mirror* 28/06/16)

The *Guardian* (27/06/16) claimed that "more than 100 reports of racist incidents" were collated in the week after the referendum. The message that Brexit and an increase in hate crimes are connected resonated and was reinforced in the post-referendum coverage of the *Guardian* and the *Mirror*.

Other two keywords examined were *feel* and *feelings*. How did people feel after the result was announced? What kind of feelings were most common in the post-Brexit Britain? The feelings that were taken into account in the pro-Remain press were mainly negative ones. Unsurprisingly, people who wanted to remain and voted for it felt *sad*, *hurt*, *scared*, *upset* and *shocked* by the result. Many felt *betrayed* by the government and by politicians who "let this happen". Feelings were described with adjectives as *bad*, *destructive*, even *apocalyptic*. Among the Remainers, members of ethnic minorities were said to feel *unsafe*, *vulnerable* and no longer *welcome* in Britain.

Other less predictable emotions seemed to be alleged regret for what could have been done differently and a supposed post-vote *guilt* and even "buyer's remorse" on the part of Brexiters:

(61) Some people are reporting feelings that they should have done more to prevent the country from leaving the EU. They feel narrow-minded for not questioning that the remain camp might not win, and they feel **guilt** for not better understanding the majority of people they share the country with. (*Guardian* 30/06/16)

(62) Among the **regretful** leave voters who spoke to the *Guardian*, some expressed shock at the ramifications of what they had meant as a protest vote. Others expressed feelings of betrayal over the leave campaign's rhetoric, the promises and the subsequent backpedaling by politicians. (*Guardian* 27/06/16)

The positive reactions of those who chose Brexit and were happy with their choice and the result were not given much coverage in the pro-Remain newspapers, with a few exceptions:

(63) Only one passerby went the other way: "Great – I've woke up English," he said, and with a look of deep joy, went on his way. (*Guardian* 26/06/16)

4. Conclusions

How much the media are able to influence both voters and politicians, directly or indirectly, is much debated and impossible to divine. The many factors that intervene between people's reception of a message from the media and the way in which they make a political decision are notoriously difficult to measure or predict (as the lamentable recent record of US and UK professional psephologists and pollsters has highlighted). However, the media certainly play a part in building and shaping political discourses and newspapers contain a historical record of these discourses.

We began this paper by revisiting Teubert's research into EU-sceptic voices, and then, with the use of corpus linguistics, and in particular CADS techniques, we gained an overview of the more recent debates on both sides, which, at the same time, provided indications of where to look for the most relevant details in the newspaper discussions. The overall lasting impression is that the referendum mechanism itself, by imposing a binary choice, herded voters who may have held many disparate and nuanced stances (as outlined in Section 2.8) into two deeply divided camps, neither of which could even begin to understand the reasoning and motivation of the other.

References

Baker, P. 2012. Acceptable bias: Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies* 9(3): 247–256. https://doi.org/10.1080/17405904.2012.688297

Brader, T. 2006. *Campaigns for Hearts and Minds. How Emotional Appeals in Political Ads Work*. Chicago IL: University of Chicago Press

Charteris-Black, J. 2013. *Analysing Political Speeches*. Houndmills: Palgrave Macmillan.

Dugalès, N. & Tucker, G. 2012. Representations of representation: European institutions in the French and British press. In *European Identity: What the Media Say*, P. Bayley & G. Williams (eds), 21–54. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199602308.003.0002

Galtung, J. & Ruge, M. 1965. The structure of foreign news. *Journal of Peace Research* 2(1): 64–91. https://doi.org/10.1177/002234336500200104

Gries, S. T. 2011. Methodological and interdisciplinary stance in Corpus Linguistics. In *Perspectives on Corpus Linguistics* [Studies in Corpus Linguistics 48], V. Viana, S. Zyngier & G. Barnbrook (eds), 81–98. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.48.06gri

Lakoff, G. & Johnson, M. 1980. *Metaphors We Live By*. Chicago IL: University of Chicago Press.

Louw, W. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology. In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 157–176. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.11lou

Marchi, A. & Taylor, C. 2009. Establishing the EU: The representation of Europe in the press in 1993 and 2005. In *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora*, A. Jucker, M. Hundt & D. Schreier (eds), 201–224. Amsterdam: Rodopi.

Mautner, G. 1995. *Only connect. Critical discourse analysis and corpus linguistics*. University of Lancaster. <http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf>

Morley, J. 2004. A sting in the tail. In *Corpora and Discourse*, A. Partington, J. Morley & L. Haarman (eds), 239–255. Bern: Peter Lang.

Musolff, A. 2017. Truth, lies and figurative scenarios: Metaphors at the heart of Brexit. *Journal of Language and Politics* 16(5): 641–657. https://doi.org/10.1075/jlp.16033.mus

Partington, A. 2009. Evaluating evaluation and some concluding reflections on CADS. In *Corpus Assisted Discourse Studies on the Iraq Conflict: Wording the War*, J. Morley & P. Bayley (eds), 261–303. London: Routledge.

Partington, A. 2012. Corpus analysis of political language. In *The Encyclopedia of Applied Linguistics*, C. Chappelle (ed.), 1–8. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0250

Partington, A., Duguid, A. & Taylor, C. 2013. *Patterns and Meanings in Discourse* [Studies in Corpus Linguistics 55]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.55

Partington, A. & Taylor, C. 2018. *The Language of Persuasion in Politics*. London: Routledge.

Rayson, P. 2012. Corpus analysis of keywords. In *The Encyclopedia of Applied Linguistics*, C. Chappelle (ed.). Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0247

Scott, M. 2008. *WordSmith Tools*, Version 5. Liverpool: Lexical Analysis Software.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

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

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

Taylor, C. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8(1): 81–113. https://doi.org/10.3366/cor.2013.0035

Teubert, W. 2001. A province of a federal superstate, ruled by an unelected bureaucracy. Keywords of the Euro-sceptic discourse in Britain. In *Attitudes Towards Europe: Language in the Unification Process*, A. Musolff, C. Good, P. Points & R. Wittlinger (eds), 45–88. Abingdon: Ashgate.

Vaghi, F. & Venuti, M. 2004. Metaphor and the Euro. In *Corpora and Discourse*, A. Partington, J. Morley & L. Haarman (eds), 369–382. Bern: Peter Lang.

We can do without these words

Investigating prescriptive attitudes to meaning in a specialised discourse

Gill Philip
University of Macerata

This chapter deals with a Style Guide written in 2013 for the British Civil Service. It included a list of words to avoid, from the difficult and vague, to the metaphorical. Viewing the Style Guide as a genuine attempt to resolve problems with the Civil Service's notoriously convoluted prose style, the study compares and contrasts the proscribed words with their use in the very documents that it was aimed to improve – online policy documents. It highlights discrepancies that are known to exist between the "proper" meanings of words and those that are found in the texts, and also reveals how comprehension problems are not just caused by lexis (metaphor in particular), but also by unusual or unexpected syntactical patternings.

1. Introduction

> Discourses are full of contributions of members proposing or resisting a change to … meaning. It is this social activity of negotiation that forces us to develop a capacity of interpretation, taking us beyond merely responding to stimuli and making us reflect on what is happening. It makes us conscious that something is going on. (Teubert 2010: 126)

In July 2013, the British Civil Service issued new guidelines for its staff on how to write online documents (*Government Digital Service Content principles*,[1] henceforth 'Style Guide'). Civil Service style-sheets are not new, nor are updates unusual. Yet this particular update caught the interest of the media and a flurry of articles appeared in the press, all picking up on the section dedicated to *Plain English*, in

1. <https://www.gov.uk/designprinciples/styleguide> (Downloaded 30 July 2013).

https://doi.org/10.1075/scl.87.06phi

particular a list of thirty-seven words to avoid, prefaced by the comment "we can do without these words". The media instantly labelled them as "banned words".

The Style Guide was published during the period colloquially known as the "silly season" – the time of year when Parliament is in recess and hard news is scarce. This gave journalists the chance to mock the Civil Service's notoriously convoluted prose style, at the same time engaging the not-unsubstantial portion of the British newspaper-reading public who hold strong prescriptive views about how language should be used. The issuing of the Style Guide was trumpeted loudly as a triumph for plain language over jargon. Both televised news and newspaper articles drew parallels with the 1980s BBC TV series *Yes, Minister*,[2] framing the publishing of the Style Guide within the stereotypical view of the Civil Service as a body whose purpose is to obfuscate and manipulate information using opaque language.

A list of words to avoid, such as that in the 2013 Style Guide, raises a number of questions for linguists. That the authors[3] of the Style Guide should have had reason to identify a few dozen words as problematic suggests that those words are used, and used "badly", to an extent that justifies their inclusion. What is not at all clear, however, is why these particular thirty-seven words were chosen. We can infer that the main purpose of the list was to draw particular attention to aspects of usage that are to be discouraged, including overly-complicated words, and "wrongly-used" words, and that this might best be done by highlighting a number of "culprits". We can also surmise that these words must be frequently and/or noticeably (mis)used within the Civil Service, since that would offer some tangible reason for their selection. Yet some doubts remain, in particular one which will be investigated during the course of this paper: to the trained eye, most of the listed words are metaphorical; or rather, the use that is being objected to is a metaphorical one. Consider the following extract which, as in two thirds of cases, comes complete with a comment explaining why the word is to be avoided:

> slimming down (processes don't diet – we are probably removing x amount of paperwork, etc.) (Style Guide § 1.5)

2. *Yes Minister* and its successor *Yes, Prime Minister* parodied the relationship between powerful civil servants and blundering government ministers. This relationship was prominently expressed through language: the civil servants were expert wordsmiths fluent in officialese, using their linguistic prowess to convince the Minister to act according to their wishes or to befuddle him, depending on their ulterior motives. Examples are available from <http://en.wikiquote.org/wiki/Yes,_Minister> (April 2015).

3. There is no authorial attribution, but the newspaper articles reported comments offered by "one of the authors", which indicates that it is a multi-authored work.

The comment makes it clear that the "undesirable" meaning is metaphorical, and this can be said of nearly all the comments in the document. Yet *slimming down*, like many of the words on the list, was not included under the separate list of metaphors. What therefore led the authors to identify some words as metaphors and other, equally metaphorical ones, merely as "misused"? And why ought any particular meanings of words, metaphorical or otherwise, be deemed objectionable? Are they really responsible for "empty, meaningless text", as the Style Guide suggests (§ 1.5), and if so, in what way?

These questions prompted the start of the study reported in the following pages, an investigation into the use and function of (mainly) metaphorical lexis in Civil Service documents, framed within the broader perspective of words and their meanings in an institutional setting. First of all I discuss style guides written for the Civil Service over the years in order to provide some background to the 2013 document and its contents. The Civil Service has a long-standing reputation for communicative ineffectiveness, hence the need to publish guides aimed at improving the clarity of written documents. Section 2 discusses these with particular relevance to the issues of "plain language" and "proper meaning" which they presuppose and which are fundamental to understanding the tone and content of the 2013 publication. Section 3 shifts the focus to the academic study of political and bureaucratic language, especially to the analysis of the use and function of metaphor within these genres. Together, these sections offer a solid platform from which to launch a systematic study of the "banned words". Section 4 addresses the data selection and analytical methods adopted for the present study.

The main focus of this paper, the less common meanings of common words, begins in Section 5. Uncommon meanings comprise collectively all those meanings which differ from the most salient, concrete meaning, popularly considered as a word's "proper meaning". This section also contains an initial analysis of the "banned words", grouping them into three broad categories, difficult words, vague words, and metaphors, mentioned at various points in the Style Guide. This section also deals with some noteworthy features of the listed words and paves the way for a case study in Section 6 which focuses on collocation and syntactic patterning in one of the better-represented semantic groupings of metaphorical "banned words" in the data: combat-related vocabulary. Section 7 addresses the role of grammar and syntax in communicating vague meanings. Starting with an analysis of the near-synonyms *commit* and *pledge*, it highlights how meaning does not only reside in the surface meaning of words, but also within syntactic constructions which communicate incipient actions but not the accomplishment of goals. This emerges as an important feature of policy documents, which are ostensibly intended to inform the public of action taken and results achieved, i.e. goal accomplishment, yet which in fact state action to be undertaken and desired

end results, i.e. work-in-progress and projected reality. A contradiction thus arises: the public expects to hear of concrete actions taken and results achieved, while the government needs to convince the public that it is taking action, and that those actions will lead to improvements. I argue in the concluding section that it is precisely because of this contradiction, and not simply due to the presence of particular words in the texts, that government-to-public written communication often seems vague and difficult to understand.

2. Style Guides: "plain language" and "proper meanings"

Despite media suggestions that this was the first ever style guide to be written ("Public sector jargon banned in *first style guide* for Government announcements" (Wright 2013), my emphasis), it was of course the first one dedicated specifically to *online* content:

> This style guide is for written content on the GOV.UK website.
> The first part of this guide is for the whole of GOV.UK. Any information on www.gov.uk (https://www.gov.uk/) should follow this guidance.
> Then there are separate sections for the 'mainstream' content (this is content for both businesses and citizens) and Inside Government content.
>
> (Style Guide, introduction)

The Civil Service has a long history of using reference works to ensure consistency across authors and topics, as well as to eradicate convoluted, opaque prose. The reference to "plain English" is inescapably related to Gowers' *The Complete Plain Words* (1954[1986]) which in turn drew heavily on Fowler's *A Dictionary of Modern English Usage* (1926[1965]), as well as a long list of works on good English style by distinguished authors including Flesch (1946), Ogden and Richards (1936), Partridge (1950), and Orwell (1946). And it is precisely Orwell's presence that is worth highlighting, because the set of prescriptive maxims that appears in one of his better-known essays, *Politics and the English Language* (1946) resonates in the Style Guide, both in its prescriptive tone and in the types of language that are highlighted, which include metaphor, long words, jargon and other expressions which are not considered to be "everyday English".

(i) Never use a metaphor, simile or other figure of speech which you are used to seeing in print.
(ii) Never use a long word when a short one will do.
(iii) If it is possible to cut a word out, always cut it out.
(iv) Never use the passive when you can use the active.

(v) Never use a foreign phrase, a scientific word or a jargon word if you can think of an everyday English equivalent.
(vi) Break any of these rules sooner than say anything outright barbarous.

(Orwell 1946: 264)

The influence of these maxims is also evident in other style guides currently circulating, including the Plain English Campaign's *How to write in plain English* (2009) and the Council of Europe's *How to Write Clearly* (no date), as well as works on plain English for the legal profession by Schiess (2003) and Kimble (1996, 2002). Linguists will notice that, to a large degree, they parallel the maxims of Quantity, Relation and Manner in Grice's cooperative principle (Grice 1975), but should readily acknowledge that the general public is not altogether familiar with Grice. Orwell, on the other hand, is far more familiar: his views on the English language are still widely-acknowledged by writers on good style (Fischer 2007), and his works still feature in the high-school syllabus in the UK. The similarities between his maxims and the structure and content of the Style Guide are impossible to miss.

At first glance, and despite the prescriptive tone, these sorts of maxims seem perfectly reasonable: writers should avoid expressions whose meaning is not immediately clear, such as clichés and formulaic phrases, as well as long, foreign or overly-specialised and technical words; and they should be concise and to the point. Yet maxims are rules, and while rules may be clear, how to implement them might not always be immediately apparent. The more helpful style guides go beyond the mere stating of rules and actually give examples of what to avoid (and why), offering suggestions as to how to transform it into something clearer. For instance, in the Council of Europe's guidelines on *How to Write Clearly*, readers are given many opportunities to compare and contrast "bad" and "better" versions of sentences. The same guide explains e.g. *why* nominalisations are cumbersome in English prose, and provides examples showing how to rephrase them effectively. The Style Guide is not so constructive. It is little more than a collection of "do's and don'ts" which leaves interpretation of the underlying reasons to the reader, and little guidance on what alternative expressions can be used instead. In the section that this chapter examines in detail (§ 1.5), the list of banned words is not presented rationally (i.e. by topic, or type of language problem), but simply in alphabetical order; and only very generic reasons are given for avoiding them. When guidance is offered, it is presented as a concession ("avoid this unless…"), and the nearest thing to a suggested alternative has to be extrapolated from those comments which indicate appropriate collocates or semantic domains of use (see Section 5) which allow the word to be used with its "proper" meaning.

Proper meanings are not at all the same thing as plain language, but the two are often conflated. This indeed seems to have happened in the Style Guide:

despite *Plain English* being the title to Section 1, the focus seems to be "proper English". This takes us into murkier waters. While there is a considerable quantity of scholarship dealing with 'Plain English' (see especially Crow 1998, Garner 1987, Kimble 1996, 2002; Schiess 2003), notions of what is "proper" slip off into the realms of folk belief, prejudice, conservatism and – in the United Kingdom, at any rate – value judgments that are strongly influenced by the prescriptivism of traditional schoolbooks. It is an area that wiser authors treat with delicacy; Gower, for instance, while prescriptive insofar as he insists on using the correct word for the meaning being expressed, has a pragmatic and open approach to new words and meanings (he dedicates two chapters of *Plain Words* to dispelling the myth that new meanings are somehow defective). The 2013 Style Guide is, in this respect, far more conservative. It resists and dismisses new meanings with school-marmish censure, taking issue primarily with new meanings of words, including nominalisations and conventionalised metaphorical uses.

"[N]ew words are not always or even usually welcomed", Richards (1936: 76) tells us in his description of the *Proper Meaning Superstition*. This is, in brief, the assumption "that words have, or should have, proper meanings which people should recognize, agree about and stick to" (Richards 1936: 72). The *Proper Meaning Superstition* "leads us to think that a shift of meaning is a flaw in discourse, a regrettable accident, instead of a virtue" (Richards 1936: 72), perhaps because the "[w]ords we meet on the street or at the supermarket threaten us less than words which creep only occasionally out of a dictionary" (Crow: 1988: 87). Plain words are simple(r) words, sometimes not up to the task of expressing complex subjects but frequently acceptable in the place of technical terms and jargon (Kimble 1996). "Proper meanings", on the other hand, are the realm of mere pedantry. While all authors on good style agree that slang terms ought not to find their way into formal and official prose, objecting to established meanings such as *key*, meaning "important", is simply pedantic. Yet it is precisely this sort of resistance to language change that defines the Style Guide's approach to words to avoid. While some of the words on the list are admittedly ugly (*incentivise* and *disincentivise*), or associated with clumsy phraseology (*dialogue*, in *engage in dialogue with* could, indeed, simply be expressed as *talk to*), few of them strike us as innovative or modish. Indeed their very normality makes us wonder why they were picked out for special mention, and why there should be such vehement objection to their use.

3. The language of politicians and politics

The academic literature on what is broadly defined as 'political discourse' is vast and ever-growing, encompassing scholars from a wide range of disciplines and

with perspectives spanning the entire political spectrum. In this section I limit the field to studies whose primary focus is on metaphorical lexis, since the striking feature of the "banned words" is that most of them are metaphorical to some extent. This is indicated by comments accompanying many of the words (see Section 5): many of these suggest acceptable collocates or semantic domains which restrict the words' "proper meaning" to a basic (i.e. non-figurative) one. Since the key identifying feature of metaphor in text is domain shift, or, less technically, the use of a word in a different context from usual (see Pragglejaz Group 2007), these suggestions imply that metaphorical meanings are to be done away with. An overview of the literature treating metaphor in political contexts is therefore called for.

As we saw in Section 2, bureaucratic texts are often considered problematic in terms of their lexical and syntactic features, and it is that mode of delivery, rather than the content as such, that attracts criticism and comment. Yet delivery and content cannot be so easily separated. The words and expressions that are used to communicate concepts obviously contribute to their meanings, but they often do so in subtle ways. Metaphor is a case in point. While metaphors are not usually thought of as "proper meanings", they may well be *common* meanings, and the ubiquity of metaphor is well known: "we cannot get through three sentences of fluid discourse without it" (Richards 1936: 92). They may or may not be used intentionally; they may or may not be interpreted as metaphors; but they are present nonetheless, and arguably shape our thinking in political discourse and elsewhere.

Political discourse has overwhelmingly – and "disproportionately", in Brougher's (2012: 145) view – focused on metaphors used by the political elite. Researchers' attention has primarily been directed towards the language produced by politicians in a range of contexts including speeches made in parliament (Chilton 2004) or at party conferences (Charteris-Black 2011), public addresses (Charteris-Black 2004, 2011; Chilton 2004, Chilton & Ilyin 1993); communication with the media in interviews and press conferences (Chilton 2004, Partington 2003, Howe 1998), as well as written texts such as party manifestos (Charteris-Black 2004). There is also considerable scholarship revolving around media representations of these same communicative situations (e.g. Johnson 2005, Howe 1998, Musolff 1998, 2004; Thornborrow 1993). From such research, we discover that political discourse is easily viewed as a strategic form of communication (Chilton 2004: 45–47), one in which the language used plays a key role in shaping the public's perception of a politician, a government, or any given policy. In short, it is persuasive. In order to gain consensus both at the individual and governmental levels, politicians have to "tell the right story" (Charteris-Black 2011: 28), and one of the many ways in which they do so is through metaphor.

> Metaphors enable you not only to communicate a complex situation with a very few words and to say a great deal more than appears on the surface, but also to capitalize on the prejudices of mankind by presenting comparisons which may not be truly analogous in some important hidden factor. (Gulick 1984: 371)

This comment, made by a (US) Civil Service insider, may come across as somewhat cynical. More measured is Charteris-Black's explanation: that metaphor makes available to its recipient "a certain mental representation that reflects a shared system of belief as to what the world is and culture-specific beliefs about mankind's place in it" (Charteris-Black 2011: 454).

The studies cited above not only tend to be overly-focused on the political elite; they also tend to treat metaphor as a more or less deliberate rhetorical device. But as Gibbs (2015: 269) reminds us, academic studies of metaphor in politics fail to take into account the ways in which metaphor is received by the audience, and thus risk overinflating its importance. Yet research from psycholinguistics offers some empirical evidence for metaphor's persuasive power. A series of studies have sought to measure the effect of metaphor on the public's comprehension and appreciation of policy (Lau & Schlesinger 2005, Landau & Keefer 2014, Landau, Keefer & Rothschild 2014, Thibodeau & Boroditsky 2013). Throughout, researchers noticed that metaphors could create "framing effects", i.e. that the consistent use of a particular metaphor in discussing one topic led members of the public not only to view that topic through the same metaphorical lens, but also to view subsequently-presented topics in the same terms. For example, if the financial crisis were couched in expressions related to traffic accidents (e.g. "The economy is *veering off course*", Landau & Keefer 2014: 470), not only did entailments from the metaphors such as agency, speed, and direction transfer onto the perception of the crisis, but they also spilled over to other topics that were introduced afterwards. While some argue that the results obtained through these studies may simply reveal that it is exposure to linguistic expressions, and not to metaphors per se, that is the key factor here (see especially Steen, Reijnierse & Burgers 2013), the evidence still strongly suggests that metaphorical lexis can subliminally influence the general public's beliefs and understanding of social and political issues. This finding can act as a counterbalance to the accusation levelled by Gibbs (2015) regarding the lack of attention paid to "everyday" metaphorical lexis within most academic studies of metaphor in political discourse. It is not only the rhetorically-employed metaphor that can shape public opinion: "everyday" metaphors would seem to have this power too. An example of this is offered by Hastings (1998: 206), who points out how the use of action verbs such as *focus*, *deploy*, *make available*, *work together*, endow the government and its agencies with "the status of deliberate, conscious actors". This sort of subliminal framing is influential. It can ensure continuing public approval of a government's actions, even though "what precisely

the work and the efforts entail and the exact nature of the objectives are frequently left unspecified" (Partington 2003: 203).

The linguistic features of politicians' language cannot, however, be simply transferred over to another genre, even though it too is closely-related to the political sphere. The language of politicians is not bureaucratic language, and has little bearing on the style and content of Civil Service documents such as those appearing on the gov.uk site. Equally – and the similarities in nomenclature notwithstanding – "public policy documents" are also a distinct genre with their own peculiarities. Public policy documents are generally understood to be the lengthy, copiously-detailed reports prepared by governments to identify problems and propose solutions through policy-making; those that in the UK context first appear as green papers, then white papers, then are converted into law after repeated rounds of discussion and consultation in the parliamentary arena. These too have attracted academic interest, but on a far smaller scale than politicians' language: analysis is usually restricted to either a single document (Hastings 1998) or, more commonly, to a single policy or socio-political issue, e.g. welfare reform (Fairclough 2000), higher education reform (Saarinen 2008), strategic urban planning for a "World City" (Flowerdew 2004), social exclusion (Koller & Davidson 2008), rather than policy as a genre in itself, or the policies of a particular administration (but see Ho (2016), on a series of policy reforms enacted by the Hong Kong Special Administrative Region).

The "online policy documents" featured in the present study are another genre entirely: they are concise texts, designed for on-screen reading, which inform the public about how governmental bodies are implementing legislation that has already been passed in parliament. Their intended readership is the lay public, not politicians and civil servants, thus the texts presuppose limited knowledge of legislative matters and terminology. Their content is restricted to the essential facts surrounding the implementation of policy: why there is a need for action, what action is to be taken, who the intended beneficiaries are. There is no room for narrative, and details and data are kept to a minimum. They are "political" only to the extent that they deal with the day-to-day implementation of the policy decisions of government. There is little scope here for persuasion and rhetoric, since they are aimed at individuals who have no power to influence the decisions taken and very little opportunity even to become directly involved in their implementation. The language is correspondingly "impoverished", hence the reputation that it has earned itself for being dull and lifeless on the one hand, jargon-ridden, cumbersome and impenetrable on the other. It is a form of officialise, otherwise known as the 'register of public administration', or simply 'administrative language' (Longe 1985, 1999). There appear not to be any published studies dealing with this particular genre, let alone any which address lexical patterning using corpus data.

4. Data and methods

The Style Guide, as noted in earlier sections, is intended for use in the writing of online content on the gov.uk website. This website is "for anyone who has an interest in how UK government policies affect them" (Style Guide § 1). On visiting the website, the user becomes immediately aware of how many topics are dealt with, from how to apply for a passport to how to pay the right amount of tax to claiming benefits. But it is the online policy documents that make up the majority of what can properly be defined "online content": it is here that the public gains direct access to information about Government policy, as opposed to directions on where to go for services, who to ask for information, downloadable versions of print documents, and so forth. In order to examine the "banned words", therefore, it was decided to focus exclusively on online policy documents. This decision also made it possible to collect data published online in a clear time-frame, since the policy documents are all dated (other content may not be). Given the time of the Style Guide's publication (mid-2013), the time-frame chosen was a full calendar year (2013) comprising the six months prior to the guide's appearance, and the six months following it. The texts were downloaded directly from https://www.gov.uk/government/policies over the period July 2013-January 2014 and saved as individual text files for use with corpus analysis software. The total size of the corpus thus obtained amounted to just over 160,000 tokens, comprising 173 texts covering the full range of policies implemented in the year 2013. Section 4.1 discusses some of the main characteristics of the texts in the corpus, while the initial analysis is presented in Section 4.2.

4.1 Online policy documents

Online policy documents are short texts, their word-count only exceeding 2,000 in exceptional circumstances. More than three quarters of the texts fall within the range 400–1,300 running words, yet a glance at Figure 1 makes it clear that it makes little sense to talk of an "average" length: there are about 15 texts per "slice" of hundred-word increments within this range. There is no clear reason for some texts to be considerably longer or shorter than others; for example, the longest text, *The health of poor people in developing countries* (22 March 2013), deals with a topic that is also discussed in several other documents in the data set, emanating from the same Ministry, and which all have a lower word-count; similarly, the shortest text (252 words), *Creating a transparent justice system* (9 April 2013), also deals with a recurrent policy issue which elsewhere requires a higher word-count.

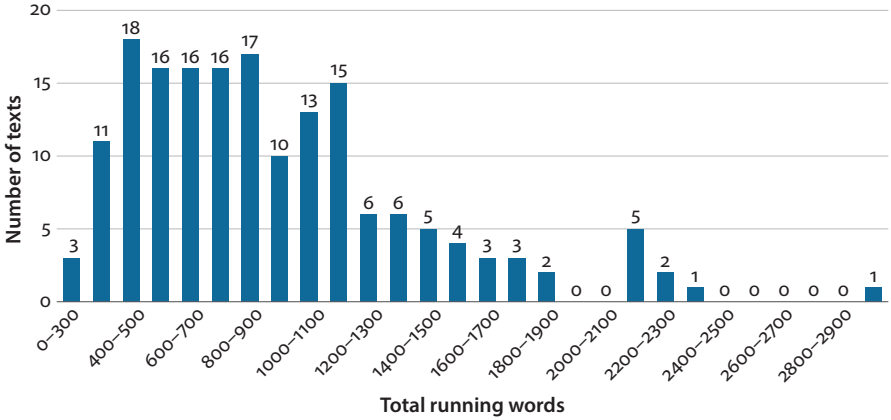

Figure 1. Distribution of texts on the basis of their length in running words (in 100-word "slices")

As mentioned in Section 3, online policy documents are intended to be the general public's first port of call in obtaining information about those government actions which might affect them. We have already noted that they are short texts. Another characteristic is their clear and simple layout: all follow a standard template. This starts with the heading *Issue*, which introduces a short section (around 100 words) in which the topic and main problems are clearly laid out (a strategy referred to in the Style Guide as "frontloading content"). *Issue* is followed by *Actions*, which introduces the main body text of the document; this part is subdivided into headed subsections where further articulation of the issue is provided, often as bullet-point lists. More complex and detailed information comes towards the end of the document, in the *Background*, which explains the rationale for the policy in more detail than can be found in under *Issue*, and also specifies *Who we're working with*. A final section, *Supporting detail*, is in effect a bibliography, containing summaries of existing legislation where relevant, plus links to related pages and downloadable pdf files.

The language used is worth commenting upon here, even before investigating the "banned words" in detail. One of the most noticeable aspects of style is the preference for present participles: all 173 document titles start with a present participle, the most frequently-used being *Making* (34), *Helping* (18), *Improving* (18), *Reducing* (18), e.g. *Making roads safer*, *Helping and supporting victims of crime*, *Improving opportunities for older people*, *Reducing drugs misuse and dependence*. This stylistic feature is also present in the texts themselves, where it is the preferred form for headings (e.g. *Prioritising mental health*, *Preventing young people from becoming drug misusers*). This conspicuous feature aside, the language looks as if it has been carefully formulated to favour short, simple words whenever possible,

and to keep sentences short – precisely as indicated in the Style Guide – although even using short, simple structures does not necessarily correspond to a reduction of formulae nor does it guarantee comprehensibility.

4.2 Investigating the "banned words"

As a first stage in analysis, the downloaded policy documents were loaded into *WordSmith Tools* 5.0 (Scott 2008), and each word on the "banned words" list was searched for in turn. Being singled out for attention in the Style Guide, it is reasonable to suppose that they should be present in a year's worth of data, but as can be appreciated from the frequencies in Table 1, column 2, many were not, and many others occurred too infrequently for any attempt at profiling to be possible. For this reason, a second search was conducted, extending the original search term to include other inflected forms. The search term used and number of hits from this second data extraction can be seen in Table 1, columns 3–4, with the different word forms and their frequencies detailed in column 5. The combined total number of hits is indicated in column 6.

Table 1. Frequency of the "banned words" in the 2013 gov.uk policy documents corpus

Word	Hits	Variant	Hits	Details of variants found	Total hits
advancing	0	*advanc**	15	1 *advances* (v); 1 *advance* (n); 1 *advances* (n); 1 *advancement*; 8 *advanced* (adj); 3 *in advance*	**15**
agenda(s)	3	*agend**:	1	1 *agendas*	**4**
collaborate	0	*collabo-rat**	13	1 *collaborated*; 8 *collaboration*; 4 *collaborative*	**13**
combating	4	*combat**	12	9 *combat* (v-inf); 2 *combat* (n-mod); 1 *non-combat*	**16**
commit	3	*commit**	70	16 *committed* (v); 1 *committing*; 27 *committed* (adj); 22 *commitment*; 4 *commitments*	**73**
countering	6	*counter**	10	10 *counter* (v-inf)	**16**
deliver	33	*deliver**	36	2 *delivers*; 8 *delivered* (v); 7 *delivering*; 19 *delivery* (n)	**69**
deploy	4	*deploy**	13	1 *deploys*; 1 *deploys*; 2 *deploying*; 8 *deployment*; 1 *deployments*	**17**
dialogue	5	*dialog**	0	–	**5**
disincen-tivise	0	*disincen-tiv**	0	–	**0**
incentivise	1	*incentivi**	0	–	**1**
drive	4	*driv**:	42	5 *driven*; 3 *driving*; 2 *drive forward*; 1 *drive up*; 1 *drive down*; 25 *driving* (n)	**46**
drive out	1		0	(see *drive*)	**1**

Table 1. (*continued*)

Word	Hits	Variant	Hits	Details of variants found	Total hits
empower	2	*empower**	4	1 *empowering*; 3 *empowerment*	**6**
facilitate	5	*facilitate**	3	1 *facilitates*; 1 *facilitated*; 1 *facilitating*	**8**
focusing	3	*focus**	31	26 *focus* (n); 2 *focus* (v); 1 *focused* (v); 2 *focused* (adj)	**34**
foster	4	*foster**	37	2 *fosters*; 1 *fostering* (v); 22 *foster* (n-mod); 10 *fostering* (n-mod); 2 *fostering* (n)	**41**
going forward	1	*forward*	0	–	**1**
impact (verb)	1	*impact**	119	1 *impacting*; 71 *impact* (n); 24 *impact* (n-mod); 23 *impacts* (n)	**120**
in order to	24	n/a	0	–	**24**
initiate	0	*initiat**	1	1 *initiated*	**1**
key [adj.]	14	*key**	(12)	12 '*Key Stage #*'	**26**
land (verb)	0	*land**	4	4 '*minimum landing size*'	**4**
leverage	0	*leverag**	1	1 *leverages*	**1**
liaise	1	*liais**	4	4 *liaison*	**5**
one-stop shop	0	*shop*	0	–	**0**
overarching	0	*overarch**	0	–	**0**
pledge	0	*pledg**	8	3 *pledged* (v); 4 *pledge* (n); 1 *pledges* (n)	**8**
progress (verb)	4	*progress**	77	1 *progressed* (v); 75 *progress* (n); 1 *progression*	**81**
promote	48	*promot**	48	29 *promoting* (n-ger); 2 *promoting* (v); 2 *promoting* (adj); 5 *promotes*; 2 *promoted* (adj); 5 *promotion*; 3 *promotions*	**96**
ring fencing	0	*ring*	0	–	**0**
robust	1	*robust**	0	–	**1**
slimming down	0	*slim**	1	1 *slimmed-down*	**1**
streamline	1	*streamlin**	2	1 *streamlined* (adj); 1 *streamlining* (v)	**3**
strengthening	12	*strengthen**	28	23 *strengthen*; 3 *strengthened* (v); 2 *strengthens*	**40**
tackling	25	*tackl**	10	7 *tackle* (v-inf); 3 *tackled*	**35**
transforming	12	*transform**	15	8 *transform*; 2 *transformed*; 5 *transformation*	**27**
utilise	0	*utili**	0	–	**0**

Part-of-speech specifications adopted: (v) verb, (v-inf) infinitive verb; (n) noun; (n-ger) gerundial noun; (n-mod) noun modifier; (adj) adjective

Even by expanding the search, a number of the words still did not occur at all (*disincentivise, one-stop shop, overarching, ring fencing*, and *utilise*) or occurred as hapax legomena (*incentivise, drive out, going forward, initiate, leverage, robust*, and *slimming down*), making it impossible to consider them for analysis. We must assume that they appear in documents circulating in the Civil Service – because for the authors to have noticed them we must assume that they are being used (and possibly abused) somewhere in the sector – but they are evidently not being used in the type of documents explicitly covered by the Style Guide.

Once the presence of the "banned words" (using the extended search terms) was checked in the corpus, concordances were extracted and the words profiled on the basis of their collocations, colligations, and semantic preferences as a way of determining their meaning in the texts (in contrast to the meanings that may be suggested in the Style Guide; see Section 5). Since the number of occurrences of the words was contained, the procedure was carried out by hand, using print-outs of the concordances. As can be appreciated from the concordance of *combat* (Figure 2), the regularities of collocation are perfectly apparent without there being any need to resort to automated procedures within the concordance software to extract collocations: in this example, the collocates *climate change*, *HIV, AIDS and malaria*, and the collocation *combat stress* are immediately apparent, as is the frequent use of *combat* as an infinitive verb (7 times). *Serious crime*, *online hate*, and *international drugs trade* can be grouped together as types of criminal

1	olice in Scotland for the purpose of	**combating**	serious crime under RIPA 2000. Given tha
2	ality, improving maternal health and	**combating**	HIV and AIDS, malaria and other diseases
3	to 2012. Case studies DFID Research:	**Combating**	climate change is a unique challenge for
4	to date. Case studies DFID Research:	**Combating**	climate change is a unique challenge for
5	nd meteorologists share knowledge to	**combat**	climate change Traditional knowledge and m
6	nd meteorologists share knowledge to	**combat**	climate change Traditional knowledge and mo
7	Programme One of the FCO's tools to	**combat**	these threats is our Counter-Proliferation
8	ing hate online This project aims to	**combat**	online hate in blogs and user-generated con
9	available online funding the 24-hour	**Combat**	Stress mental health helpline for veterans
10	twork, the Royal British Legion and	**Combat**	Stress to make sure that more veterans know
11	health care MDGs in focus – MDG 6:	**Combat**	HIV, AIDS, malaria and other diseases Find
12	DFID Research: Straight Talking to	**combat**	HIV among young African women A project in
13	Supporting rural health workers to	**combat**	malaria Giving women a choice in Malawi How
14	Parliament to support equality and	**combat**	discrimination. We contribute to the develo
15	will start a new, and smaller, non-	**combat**	mission in Afghanistan based on training, a
16	anti-terrorism measures to helping	**combat**	the international drugs trade. Casualty and

Figure 2. *Combat** in the 2013 gov.uk policy documents corpus

behaviour, forming a third semantic preference grouping, in addition to environmental issues, and illness, which are represented by the recurring collocates already listed.

To test whether or not the "banned words" in the gov.uk policy documents corpus behave in unusual or surprising ways, the same terms were extracted from the *British National Corpus*. In this way, their lexical profiles could be compared against a general reference corpus of British English, and any important differences noted. Some of these comparisons are discussed in Section 6. Before comparing the gov.uk policy documents and general English, however, it is time to take a closer look at the words and their purported meanings, comparing them to the corpus data.

5. Pretexts for banning words

Complementary to the corpus-linguistic profiling of the words is an analysis of how the Style Guide presents them. It relies on a peculiar hybrid of decontextualisation and hypothetical contextualisation: the words are simply presented as an alphabetical list, and many are accompanied by comments suggesting appropriate collocates or semantic domains of use. The user is expected to understand from minimal indications what uses and meanings of the words are acceptable and, by exclusion, which are not; and in the absence of any comment, it is difficult to understand not only what to use instead, but why the words are to be avoided in the first place. In this section, a provisional categorisation of the words is offered, drawing on information gleaned from the Style Guide itself (Section 5.1). After this, their permitted meanings – as understood from the lexical and semantic information provided in the comments – will be indicated (Section 5.2), and data from the corpus will be used to highlight some areas where the Style Guide's prescriptivism comes unstuck.

5.1 Difficult words, vague words, and metaphors

On the basis of the topics dealt with in the sections immediately preceding the list, and thanks to the parenthetical comments which indicate the circumstances in which they are permitted, the "banned" words and expressions can be grouped into three general categories. These represent areas of language use which the Style Guide authors deem problematic: "difficult words", "vague words", and "metaphors". The difficult words are mainly words of Latin origin, all which have readily-available "simpler" alternant forms. Vague words require paraphrasing to make their meaning clear. Metaphors are all the words which are, in truth, *not* to

be used metaphorically – as indicated in the accompanying comments in the Style Guide, reproduced in Tables 2–4.

The preamble to Section 1.5 of the Style Guide starts by stating "Don't use long or formal words when easy or short ones will do. Use *buy* instead of *purchase*, *help* instead of *assist*, *about* instead of *approximately* and *like* instead of *such as*." Here we see words of Latinate origin substituted by more everyday terms. Two words on the list, *collaborate* and *dialogue*, are presented together with a suggested "plain English" synonym: *use "working with"* and *we speak to people* respectively. To these we can add a number of the uncommented words, all Latinate expressions which have more immediately-comprehensible alternatives: *disincentivise*, *incentivise*, *initiate*, *liaise*, and *utilise*. We can also assign to this list of "difficult words" the polyword *in order to*, which is described as "superfluous – don't use it" (and – inexplicably – is included under the sub-list of metaphors to avoid), as well as *overarching*. Table 2 shows the group of "difficult" words.

Table 2. "Difficult" words mentioned in the Style Guide

"Difficult" words	**Commentary**
collaborate	use 'working with'
dialogue	we speak to people
disincentivise (and *incentivise*)	–
initiate	–
liaise	–
utilise	–
in order to	superfluous – don't use it
overarching	–

A fine line separates "difficult words" from "vague words"; the criterion referred to was the suggestion not of an alternative expression but rather the invitation to be more explicit. Words that fall clearly into this category, because of the comments, are *commit/pledge*, *facilitate*, *progress* (verb), and *transforming*. Amongst the uncommented terms, *advancing*, *combating*, *countering*, *empower*, *focusing*, *impact* (verb), and *robust* also belong here; see Table 3 for the complete list. Like "difficult" words, vague words too can be expressed in more simple language, e.g. *going forward*, *fighting*, instead of *progressing*, *combating*, but even when paraphrased, their meaning remains vague. Indeed *going forward* appears on the list of metaphors to avoid (see Table 4), which suggests that simplicity of word-form is not enough to resolve some problems of meaning. Vague words, in brief, require paraphrasing, not substitution, which may of course be easier said than done considering that as well as avoiding these words, document writers are also expected to avoid

"unwieldy grammar" and "long sentences with complicated sub-clauses" (Style Guide § 1.3).

Table 3. "Vague" words mentioned in the style guide

"Vague" words	Commentary
commit/pledge	we need to be more specific – we're either doing something or we're not
facilitate	instead, say something specific about how you are helping
progress	as a verb – what are you actually doing?
transforming	what are you actually doing to change it?
advancing	–
combating	–
countering	–
empower	–
focusing	–
impact (verb)	–
robust	–

We understand that the remaining words are metaphors from the commentaries suggesting suitable collocates and/or semantic domains – implying that the words are being used outside these strictures. These commentaries variously suggest collocates, semantic domains, or a combination of the two, and the effect is similarly independent of the presentation: collocates help to fix the basic, or "literal", meaning in its most immediate, concrete sense, while semantic domains indicate the sphere of reference. *Leverage*, for instance, is permitted "in the financial sense", but not in any extensions of it, and *foster* should only be used of children (i.e. not as *fostering creativity* or other such senses). Table 4 lists the metaphors and their comments, with collocates in italic font and semantic domains underlined.

What is interesting about this third group is how many words are primarily metaphorical. The Style Guide lists six expressions as metaphors, one of which (*in order to*) is definitely misplaced. To the remaining five, we can add a further ten on the basis of the commentaries alone, plus *streamline* (here taken to be a synonym of *slimming down*), and at least seven of the "difficult" and "vague" words (*advancing*, *combating*, *countering*, *empower*, *focusing*, *overarching*, and *robust*). Strictly speaking, only four of the 37 entries are not potentially metaphorical: *in order to*, *incentivise /disincentivise*, and *utilise*. Why this should matter is that, in the first place, it takes us back to the earlier discussion about plain language and "proper" meanings, because if the authors explicitly state that metaphors are to be avoided as well as listing a series of other words which are not to be used non-literally,

Table 4. Metaphors with suggested suitable collocates and domains

Metaphor	Commentary with suggested *collocates* and domains
agenda	unless it is for a *meeting*
deliver	*pizzas, post* and services are delivered – not abstract concepts like '*improvements*' or '*priorities*'
deploy	unless it is military or software
drive	you can only drive vehicles; not *schemes* or *people*
drive out	unless it is *cattle*
foster	unless it is *children*
going forward	unlikely we are giving travel directions
key	unless it unlocks something
land (verb)	only use if you are talking about aircraft
leverage	unless in the financial sense
one-stop shop	we are government, not a retail outlet
promote	unless you are talking about an *ad campaign* or some other marketing promotion
slimming down	processes don't diet
streamline	–
strengthening	unless it's strengthening *bridges* or other structures
tackling	unless it is *rugby*, *football* or some other sport

metaphor is being cast as something that is difficult to understand, and/or as responsible for vague meaning, and/or as something that can, and should, be eliminated (Style Guide § 1.5). In other words, what is being presented is the traditional (Aristotelian) view of metaphor as a decorative device which can be substituted by a literal paraphrase (see Black 1962: 25–47, Deignan 2005: 2–4). It also brings into play the idea of metaphor's persuasive power in political contexts, and its contribution to the subliminal framing of policy, which are somewhat at odds with the Style Guide's authorial stance demanding straightforward, clear, jargon-free, "everyday" language in communication with the general public.

5.2 Always avoid metaphors

Grouping the listed words into "difficult words", "vague words" and metaphors allows us to extrapolate what the overall purpose of this section of the Style Guide is, but takes us only so far. The main problem with presenting an alphabetical wordlist is not understanding why some words are present and others absent, but

to determine what meanings are intended. Understanding decontextualised words is very much a matter of inference. Rather than communicating meanings, they only have 'meaning potential' (Hanks 2013: 87–88), and as such, they appeal to the folk-linguistic belief that words have "proper" meanings. This belief is reflected in the Style Guide, and its shortcomings come to the fore when the "banned words" are looked at in context: it utterly fails to take into account contextual meanings and terminology.

Most of the words on the list that are commented upon have been singled out in order to suppress their use in semantic domains or collocations which shift their meaning away from the basic, i.e. a meaning that is more concrete, related to bodily action, more precise, and/or historically older (Pragglejaz Group 2007: 3), though "not necessarily the most frequent meanings of the lexical unit" (Pragglejaz Group 2007: 3). Two problems can be identified here. The first is that there may be a mismatch between what is familiar because frequent, and what is familiar because basic. The second is that there may be literal meanings which are not basic but impossible to substitute or paraphrase precisely because they are literal. Consider *land*, which the Style Guide states is only to be used in the collocation *land + aircraft*. We might infer from this that figurative meanings such as *land + job/contract* are to be avoided. In the gov.uk policy documents corpus, however, all four occurrences of *landing* (the only form found) are part of the technical term *minimum landing sizes* (see Figure 3). This refers to the size of fish taken off boats when they unload at port, i.e. when they *land fish*. In the context of fisheries policies, this use of *land* is literal, but it is not basic. With this example we see evidence of the Style Guide authors making an over-generalisation about meaning which could have been avoided had they explained properly what they wanted authors to avoid (as happens in other style guides) rather than rely on prescriptive and impressionistic views of meanings which are not relevant to the contexts that the policy documents deal with.

1 s. If the fish caught are below the minimum **landing** sizes, they must be discarded. An e
2 ing small juvenile fish, called the minimum **landing** sizes technical rules. However, EU
3 re of quota species below the legal minimum **landing** sizes and too small to land. Market
4 seas around the UK. Under the legal minimum **landing** size There are rules to stop fisher

Figure 3. *Land** in the 2013 gov.uk policy documents corpus

Comparable results are obtained with most of the other words on the list whose ideal meaning can be understood from the comments (i.e. those listed in Table 4). For example, *deliver* already collocates with *services* (14) and other words which can be grouped into the same semantic preference, e.g. *plans* (3), *projects* (10), *programmes* (17), in accordance with the Style Guide's instructions. In this instance

the comments conflate two distinct meanings: to deliver something somewhere, the most frequent and also the basic meaning,[4] accounts for the collocates *pizzas* and *post* in the Style Guide commentary, while *services* belongs with the second-most frequent meaning, "to provide a service" (sense 5 of 11 in MEDAL). The imposition that only concrete services can be delivered while abstract ones like *improvements* or *priorities* cannot is likely related to an earlier comment (regarding *commit/pledge*) to be more specific about the action being taken; but the relevance of such a meaning to Government documents is tenuous and, unsurprisingly, not attested in the data. Those collocates which do occur as things which are created and made accessible to the public, such as *energy* (4), *jobs* (8), *homes* (8), and *housing* (6), also belong properly under this second meaning because they are not transportable objects. Yet again, the Style Guide comment is shown to be at odds with the types of information that these documents have to communicate.

The Style Guide directions often seem superfluous. *Foster* is used with reference to re-homing children in 37 of the 41 occurrences of the lemma in the data. *Deploy* is indeed mainly used with reference to military personnel. And while only one of the four occurrences of *agenda* refers to the desired "meetings" sense, members of the public at large are arguably no more likely to be familiar with meeting agendas than they are with the political issues sense that the other occurrences refer to and which is more frequent in British English: both COBUILD and MEDAL list the "meetings" sense at the end of their entries, confirming that it is neither frequent nor basic.

What the Style Guide authors believe to be the "proper meaning" of the listed words and the meanings that emerge in the data often brings the contrasting notions of basic and frequent meaning to the fore. The suggestion made is always to eschew the more recent meaning, even if it is established as in the case of adjectival *key*. Elsewhere, specifically for those words that were left uncommented, it is the data itself which suggests details not articulated in the Style Guide, details which hint at uses of language which are unusual, not-quite-established, and possibly for this reason to be done away with. For example, two of the five instances of *progress* (verb) occur in a transitive construction rather than one of the more usual constructions, i.e. the intransitive or the middle-voice. In other words, instead of the intransitive *people progress* (Figure 4, lines 3–4), or the middle-voice *negotiations*

4. Statements relating to frequent senses draw on the *COBUILD dictionary on CD-ROM* v 1.0 (Sinclair ed. 2001), since it lists word forms strictly in order of attested frequency of use in a large corpus of general English (the 329 million-word *Bank of English*); other corpus-based dictionaries list the more basic sense first, regardless of frequency: basic meanings are therefore determined with reference to the *Macmillan English Dictionary for Advanced Learners* (MEDAL; Rundell ed. 2002), the preferred source for basic meanings in current metaphor identification protocols (e.g. Pragglejaz group 2007, Steen et al. 2010).

have progressed (line 5) – meaning that people have caused *negotiations* to *progress* but expressing the action as if the grammatical theme of the verb were in fact the agent (see Radden and Dirven (2007: 289–291) on the middle-voice construction in English) we are presented with *a programme of government support* and *government* as agents of the transitive verb *progress* (lines 1–2). In the absence of any other noteworthy features in the context of this verb, it may well be this peculiar transitivity pattern which caught the attention of the Style Guide authors.

1	a programme of government support to help	**progress**	development on large-scale sites inc
2	ent is working closely with all states to	**progress**	these actions across all 3 'pillars'
3	gh the Business Compact We'll help people	**progress**	in the labour market and ensure they
4	stem to help more people to move into and	**progress**	in work, while supporting the most
5	otiations to mitigate climate change have	**progressed**	to date. Case studies DFID Researc

Figure 4. *Progress** in the 2013 gov.uk policy documents corpus

There may also be incipient trends like this one in the use of other uncommented words: intuitively, *advancing* and *impact* seem ideal candidates for investigation, but far more data would have to be collected for this to be possible. Indeed, those words which the authors were unable to comment upon appear to be problematic for writers too, since they are already doing without them.

6. Metaphors in administrative prose

Assigning the "banned words" to the three main categories dealt with in the Style Guide makes it possible to illustrate the areas that seem to be problematic with concrete examples. It can be observed that metaphorically-used lexis is problematic, but we have also seen that the suggested meanings are not necessarily any more accessible or even relevant to any context that might interest the Civil Service. To state that "pizzas, post and services are delivered" when the Civil Service is not a food outlet, post office or manufacturer of goods, is unlikely to have much effect on how *deliver* is used in online policy document texts. The comment misses the mark entirely. Another division of the data is called for; one which allows us to differentiate between words which seem to be saying the same thing, and which can shed light on the choice of particular words for the communication of particular meanings in particular contexts.

In a disparate alphabetically-sorted list, regularities do not stand out, but many of the words can be grouped together loosely into semantic domains. It should be stressed that these are not intended as metaphorical source domains, but are simply a lexicographic sort on the basis of the words' basic meanings. The

purpose is to impose some order to the analytical procedure, making it possible to compare and contrast words which have a degree of semantic affinity. The rest of this section focuses on the first of the groupings shown in Table 5, 'Fighting and defending'.

Table 5. Semantic groupings of the "banned words"

Semantic area	Words
Fighting and defending	*combating, countering, deploy, tackling*
Forward movement	*advancing, drive, going forward, progress*
Power and strength	*empower, impact, leverage, robust, strengthening*
Promises	*commit, deliver, initiate, pledge*
Reduction and elimination	*drive out, ring fencing, slimming down, streamline*
Support and encouragement	*(dis)incentivise, foster, incentivise, facilitate, promote*
Working together	*collaborate, dialogue, liaise*
Miscellaneous	*agenda, focusing, key, land, overarching, transforming, utilize, in order to, one-stop shop*

6.1 A closer look at metaphor: 'fighting and defending'

The 'fighting and defending' category groups together *combating*, *countering*, *tackling*, and *deploy*, to which I added the superordinate *fight* for comparison purposes, since it is a "short, easy word" of the sort that the Style Guide urges its readers to adopt in preference to the "banned words". In the year's data from gov.uk, *fight* occurs 21 times, *combat* 14 (of which 4 *combating*), *counter* 16 (of which 6 *combating*), *tackle* 35 (of which 25 *tackling*) and *deploy* 17.[5] All except *deploy* are used in two main contexts: disease and crime (including alcohol-related anti-social behaviour, hate crime, as well as slightly lesser evils such as discrimination and stigma). Both of these semantic domains (particularly crime) are known to be associated with war metaphors in English (Lakoff & Johnson 1980, Steinert 2003, Charteris-Black 2004, Semino 2008) and crime is known to be framed conceptually, in some contexts, as a disease affecting society (Lakoff & Johnson 1980), making these domains interrelated conceptually, at least some of the time.

The war on crime has become a staple metaphor in political discourse (Howe 1988, Mio 1997, Steinert 2003, Semino 2008). As Semino explains,

5. Here and in the following discussion, the numbers of occurrences refer to the lemma unless otherwise stated.

> WAR metaphors are often used in relation to particularly serious and intractable problems, and to the initiatives that are developed in order to solve them. […] Metaphors such as these emphasize the gravity and urgency of the problem in question, and the seriousness of the effort that is being made to solve it.
>
> (Semino 2008: 100)

But metaphors highlight and hide in equal measure: by emphasising one particular facet of meaning, the others are necessarily downplayed. The issues referred to by WAR vocabulary may indeed be serious, but the finer details are left for the recipient to infer. Who, for instance, are the participants within metaphors of conflict, and what is their status: aggressor or defender? Are they equally "strong" and capable of winning the fight, or is one side weaker than the other? Does the aggressor have a strategy or purpose, and can it reasonably be counteracted by the defender? What weapons are employed, and what can be used as defence from them? Such details are generally not explored very far within cognitive views of metaphor because they are necessarily bound by contextual constraints, as the analysis reported in this section will demonstrate.

6.2 Collocational specialisation: 'crime', 'disease' and other collocates of 'fighting and defending'

Despite the overlap of DISEASE and CRIME as general semantic domains associated with *fight*, *combat*, *counter*, and *tackle*, each of the terms has its own specialisation of meaning and subfield of reference. Consistent with its role as superordinate, *fight* is the least specialised. It collocates with another superordinate, *crime*, but with no particular sub-class of crime. In contrast, *combat* collocates with the specific crime of discrimination and racism (particularly anti-semitism and anti-Muslim hate crime). BNC data also provides evidence of this collocational specialisation, alongside others, but considering the different periods in which the corpora were compiled, the absence of other forms of discrimination in gov.uk (e.g. sexism, present in BNC) is not surprising: religious intolerance poses the greatest challenge to law and order in the present socio-political climate and is therefore at the heart of policy-making decisions. In the BNC, *combat* also collocates with *crime* (43), *violence* (10), *terrorism* (22), and *threat* (15 – usually with reference to terrorism). Also present in significant proportions are *drugs* and *problem*, both at around 5% of all occurrences of the lemma. The *problems* identified are social, such as *unemployment* (17), *truancy* (5) and *non-attendance* [at school], and *poverty* (14), which are not necessarily crimes as such, but fall into an intermediate category of situations which are detrimental to the harmonious workings of society, and may potentially trigger criminal behaviour.

The crimes that *counter* collocates with are *threat* and *terror(ism)/terrorist(s)* which, like religious discrimination and hate crime, represent a topical problem which we should expect to find in the 2013 gov.uk data. This specialisation is peculiar, not only because it is very restricted in range, but especially because it is used to refer to hypothetical or potentially undesirable situations – i.e. not currently in existence – whose manifestation is to be prevented. This stands in marked contrast to *fight* and *combat* which are used to talk about undesirable situations existing in the present which are to be controlled, if not eradicated completely. The general norms of usage as attested in the BNC are a little less specific: while *threats* is present amongst the collocates (10 occurrences), it is mainly the effects of *growing* (9) or *increasing* (5) states of affairs – particularly verbal and physical aggression – that are the focus of *counter* in general English. That the Style Guide picked up on *combat* may, therefore, be due to its restricted collocational behaviour in policy texts, as compared to "normal" use in the language at large.

The crimes that gov.uk policies intend to *tackle* fall into two broad categories. The first is discrimination and intimidation, exemplified by collocates including *extremism and intolerance* (3), *hate* [*crime*] (3) and *antisocial behaviour* generally, including *bullying*. Also present here is white-collar crime, namely *tax avoidance* (2) and *evasion* (2), and *organised crime and corruption* (1). The themes reflect those found in the BNC, which itself has no particular preference for any one collocate. In both these categories, the "weapons" are abstract – words and figures – rather than guns, knives and bombs (*knife crime*, the only weapon-related crime mentioned, occurs just once). Such crimes are *tackled* through the passing and tighter implementation of laws, both of which are things that policy-makers can control. And thus, metaphorical *tackle* has an important role to play: it helps persuade the populace that the Government is effective, not merely *defending*, *combating* or *countering* crime, but winning the battle.

Fight and *combat* both collocate with a range of preventable endemic diseases (infections and viruses) afflicting developing countries, particularly *malaria*, *cholera*, *HIV/AIDS*, and *sleeping sickness*. Although endemic, all of the diseases mentioned are preventable, and most are treatable. Fighting these diseases is therefore primarily a matter for education and policy (hence their appearance in policy documents): their spread is to be controlled and their manifestation prevented. Treatment – the solution to the problem – is never mentioned.

Disease is the most prominent semantic field for *tackle*, in two distinct but related contexts found mainly in foreign policy documents. First there is the generic area of *poverty and vulnerability* (3), and poverty-related problems of *malnutrition* (2) and *hunger*. More specific, however, is its association with waterborne diseases, typically *malaria* (2) and *cholera* (3); or simply the general *threat of disease* (2), both with reference to the aftermath of the *Pakistan floods*, that is a consequence

of a *lack of water and sanitation facilities* (2). In other words, illnesses that can be prevented and treated by foreign aid efforts are problems which government policy can address and take control over. So too are the two domestic health issues mentioned, *binge drinking*, and *obesity*, because legislation can be passed and implemented.

In addition to actual diseases, *tackle* collocates in general English with a range of "social ills", i.e. metaphorical diseases which afflict society. *Problem(s)*, *issue(s*), and *question(s)* together account for 22% of the collocates of *tackle* in the BNC, and a wide range of social problems are also mentioned, from *unemployment* (29) to *homelessness* (13), *drugs* (5), and *poverty* (8). All of these are (still) current political issues, matters which are at the centre of domestic policy-making. *Tackling* a problem entails the notion of having the upper hand, and indicates a greater chance of success than that suggested by defensive vocabulary. It suggests a concerted effort and positive results, both of which are viewed more favourably than *combating* and *countering* such problems would be. Irrespective of whether one wishes to reclassify *tackle* as a SPORTS metaphor rather than a WAR one, there is an important rhetorical device at work when defensive vocabulary is supplanted by the offensive *tackle*. It may not be clear to readers of the policy documents that this is present under the surface text, but the Style Guide authors might just want to reconsider their instruction to do away with non-literal *tackle*: it seems to serve a politically-advantageous communicative purpose.

As well as CRIME and DISEASE, green issues also feature prominently in the collocates of the 'fighting and defending' terms, particularly with *deploy* and with *combat*. In fact, when *deploy* is not used in its literal (military) sense, it collocates with *climate change* and similar terms, but unlike the other words in this group, with neither CRIME nor DISEASE.

In gov.uk, we can find *combat* collocating with *climate change* in 4 of the 16 occurrences, reflecting a long-standing trend seen also in the BNC, where *combat* collocates with *global warming* (14) and *greenhouse effect* (9). Like crime and violence, and endemic diseases, this is an area where government policy can play a part, but only to the extent that it can limit the damage already done and try to prevent the situation from worsening. In other words, there can be no pre-emptive action (and so it should not come as a surprise that *counter* does not collocate with green issues) and no real control (*tackle* does not collocate with green issues in gov.uk, although BNC data informs us that it does so elsewhere, with both *environment* (14), and *pollution* (22) being recurring collocates).

Deploy occurs 17 times (9 as *deployment*) in 12 separate texts covering the period from 21 February to 12 November 2013. Ironically, perhaps, all but one of the occurrences prior to the publication of the Style Guide use *deploy* with the approved literal meaning, i.e. with reference to troops and medical personnel

in military or peacekeeping contexts, while 6 of the 8 occurrences appearing in the second half of the year express the more abstract sense which can be roughly glossed as 'implement', all with reference to renewable energy and pollution-reducing techniques and technology. This contrast in use can be justified in part by the types of policies that were being published – in the first semester the policies focus on foreign human rights and peacekeeping issues, while later in the year the focus was on pollution and climate issues. It is also true that it reflects a discourse on renewable energy found elsewhere in society at large. However, in the specific case of *deploy*, what is interesting is not so much the WAR metaphor as such, but the vagueness that it introduces: in the words of the Style Guide, "what are you actually doing?". With *deploy*, an impressionistic meaning comes over, derived from the military connotation (e.g. *action, efficiency, strategic planning*, etc.) but these facets of meaning are implied, not stated. Indeed, by not actually saying what is being done, but only suggesting, a convenient amount of manoeuvring space is left for government Ministers in case of heavy questioning by opposition parties or the media. This is an example of how meaningless text can be created out of meaningful words.

To sum up, what this restricted range of collocates reinforces is the underlying metaphorical implications associated with the words: government policy is set up to *fight* general aggressors, *combat* more specific ones, *counter* (pre-empt) threats of future aggression, and *tackle* (i.e. fight successfully against) those dangers which we are able to control. But the lexical collocates offer only a partial picture of this. The rest is revealed in the syntactical patternings which each of these verbs prefer.

6.3 Syntactical specialisation

It is easy to underestimate the contribution that syntax makes to meaning, especially in a context such as this where the focus is so strongly aimed at single words. But what emerged from the analysis of the "banned words" was not only the collocational specialisation of each term, but also the fixity of each one's syntactical patterning. The patterning reinforces the meanings already discussed in Section 6.2, sometimes in unexpected ways.

We saw that the superordinate term, *fight*, attracted the superordinate *crime*, and can broadly generalise from this that its use is non-specific (perhaps vague). The syntax confirms this. *Fight* occurs in simple structures, most typically *to fight…* or *the fight against….* The undesired situation, whether crime or disease, already exists and must be controlled. But precisely how, when, and by whom, we do not know. In contrast, and despite its similar collocational preferences, *combat* appears to be one step removed from actual engagement with the "enemy". Its most typical patterning indicates intention *for the purpose of combating…* rather than actual

intervention. So not only do we not know how, when or by whom; we do not even know for sure *if* the combat will take place. This is somewhat alarming given that the situations are present and in several cases observed to be increasing or worsening. What is worth noticing, however, is that there is a subtle deviation here from the patternings found in general English. In the BNC, *combat* (932) is used of undesired situations, often growing, which need to be controlled or eradicated, but it primarily expresses methods and strategies being adopted, such as *as a means/way of combating*, or *help(s) combat*. While success is not guaranteed here either, at least there is the sense of something actually being done!

The collocates of *counter* are not things which exist in the present but are perceived as threats which will manifest themselves in the immediate future; from this alone we surmise that *counter* refers to prevention rather than intervention. This interpretation is bolstered by the typical syntactical patterning which features the infinitive of purpose, with or without (another "banned" expression) *in order to*. This means that we are dealing with a declaration of intention to act rather than an account of action actually taken. And this marks, again, a contrast with the BNC data. *Counter* is used of existing threats and aggression, but it is clear that such attempts are not really expected to succeed: consider that the 'VB *to counter* NN' construction with one of only six verbs (*attempt*, *design*, *intend*, *need*, *seek* or *try*) accounts for about 10% of all occurrences of transitive *counter* in the BNC. At least these are attempts, however, not mere intentions.

In both government policy documents and in the general language, *tackle* is used when the problems to be resolved are socio-economic or environmental. What is interesting is that in both corpora, *tackle* fails to live up to the expectation of representing offensive action. In gov.uk, *tackle* is overwhelmingly found as a complement of *help*, i.e. *help to tackle…*. This catenative construction (Aarts, Chalker & Weiner 2014: 60–61) shifts the focus of the verb *tackle* onto its matrix verb *help*, so the action is no longer *tackling* as such, but merely helping to do so. Its mood is therefore conative (Levin 1993), i.e. "an attempt to do something, rather than a successful action in doing it" (Hanks 2013: 205). The modality in the BNC data is of a more familiar type, featuring the deontic modals *will*, *can*, *need*, *should* and *must* (i.e. *problems* have to be *tackled*); but there is additionally some evidence of conative with *attempt to* and *try to*. These parallel the catenative *help*-construction in the gov.uk policy documents and confirm that even in general English, the non-sporting sense of *tackling* it is not resultative. It does not refer to the accomplishment of a goal but instead to incipient action.

6.4 Summary

The conceptual metaphor of WAR sits well with a perspective of Government policies being a form of strategic intervention in a "battle". This kind of view is promoted by policy-makers because framing an issue as a "war" makes it easier to identify "enemies" (crime, disease, pollution, and those responsible for its existence and escalation) and "defenders" (the Government is taking preventative action). Reducing the finer complexities of crime, disease, or pollution to a simplistic "us versus them" view makes issues easily accessible to the general public, stimulates emotional engagement, and, if done skilfully, brings people round to the government's "side". This in turn means votes and the promise of future power. But the implications carried by the metaphorical lexis are not entirely supported by the structures in which they appear. The mood throughout the gov.uk data is conative, no successful completion of the action being foreseen. Intentions are stated, but it is not intentions that need to be delivered. Intentions cannot fail, they can only be said not to have succeeded *yet*; and it is upon this slender thread of promised action but deferred implementation that so much political persuasion hangs.

7. Phraseological environments, "wrong" meanings and vague language

It has become something of a commonplace in corpus linguistics to affirm that it is the phrasal environment of a word that fixes its meaning. Equally, it is well known that the frequent use of particular phraseologies can lead to delexicalisation, i.e. a "loss of identity" of individual words relative to the form and function of the phrase of which they form part (see Philip 2011). In the gov.uk data, something peculiar happens – at least insofar as the metaphorical lexis is concerned. Here, the words seem to have meaning, but that meaning is subverted by the phraseology. Instead of words merging into meaningful phrases with relatively meaningless components, the words in themselves appear to be meaningful but the phrases they contribute to are not. In other words, we do not find ourselves dealing with an empty lexicon (Sinclair 1996), but rather with an empty text.

Now, this apparent contradiction has to be interpreted with reference to the context in which these words are being used. Most language users use language to do purposeful things. We ask for things, ask people to do things for us, give information, and so on. The purpose of Policy documents is ostensibly to inform the public about what the Government is doing, to give them access to information and in some cases to invite them to participate. That is the official line. The more sceptical among us would say that the purpose of such documents is not

primarily informative, because while they inform the public about the government's intentions, pledges and promises, they do not inform the public of results, achievements, successes and failures.

Evidence of such "promises to act but no promise of results" comes from *commit* and *pledge* – presented side-by-side in the Style Guide and accompanied by the comment "we need to be more specific – we're either doing something or we're not". The problem with commitments, pledges, and other sorts of promises, is that their semantic reference point is restricted to beginnings; they express intention to act in the future. They are not about doing things, and they are certainly not about delivering results. In other words, they are conative, just like the verbs complementing *help*, *try*, or *attempt* that were discussed in Section 6.3. Even when found in the past tense, *committed* only refers to the moment of promising to act in the future, not to the realisation of the promised action. In the extract below, for example, the publication discusses a past *commitment* which is being *renewed* in recognition of the fact that its accomplishment is *unlikely*. Note that no results are documented – we learn only that "some local progress has been made" – and note that no target dates are set for the renewed commitment.

> We have **committed** to helping reduce hunger around the world as part of the Millennium Development Goals (MDGs), *a series of targets agreed at the United Nations in 2000*. Target 1C of the MDGs aims to halve the proportion of people who suffer from hunger between 1990 and 2015.
> *While some local progress has been made, this target is unlikely to be met* around the world. Rising food prices – fuelled by growing populations, extreme weather and the global financial crisis – meant hunger levels spiked in 2009.
> The UK *renewed its* **commitment** to reducing malnutrition in particular by joining the global 'Scaling Up Nutrition' movement of 33 developing countries.
>
> (*Reducing hunger and malnutrition in developing countries*, 4 October 2013, emphasis added)

One of the peculiar features of *commit* in the gov.uk data is its colligation with non-finite verbs and gerundial nouns. Non-finite forms have no temporal reference, which further reinforces the conative mood. It is therefore significant that the structure '*commit** + *to* + *–ing*' occurs approximately four times more frequently in the gov.uk data than in the BNC, not only because the conative mood is ever-present, but because this syntactic specialisation is not typical of the phraseology of *commit* in the language at large. Additionally, seven of the 43 occurrences of this structure are realised within the extended lexicogrammatical frame '*commit** + *to* + *working with* + [group] + *to*-INF':

> The government is determined to do all it can to ensure more new homes are delivered quickly. We're **committed** *to working in partnership* with local authorities

> and other partners *to accelerate* the delivery of large housing schemes.
> (*Increasing the number of available homes*, 20 August 2013, emphasis added)

The same lexicogrammatical frame occurs only ten times in the BNC, i.e. in only 0.17% occurrences of *commit**, compared to 9.59% of the same form in gov.uk.

Emerging here amongst the significant colligates of *commit* is the adjuvant (*helping* to do something), which has already been discussed in the 'fighting and defending' case study (Section 6). The interaction of this with the conative mood further emphasises the lack of actual activity, shifting the focus instead onto *intention* to act by *contributing* to the resolution of problems. Five of the "banned words" also realise aspects of this meaning: *incentivise* (1) and *promote* (92) which both suggest 'making things possible', and *empower* (6), *facilitate* (8), and *foster* (4) which suggest 'giving people the opportunity to become involved in the process'. But rather than criticising this use of language, we have to remember the contradiction inherent in policy documents mentioned at the start of this Section, namely that their authors are charged with the task of explaining what the government is doing when in fact all they can do is say what the government intends to do.

8. A final word about the words

It is difficult to say exactly how policy document writers are to report action ("we need to be more specific – we're either doing something or we're not"; "what are you actually doing?") when the "actions" are in fact "intentions". The "actions" discussed in policy documents are primarily things that the public expects the government not merely to do, but to complete successfully, i.e. the type of verb known as "accomplishments" within 'Aktionsart' (Bache 1995, Rothstein 2004). Accomplishments focus meaning on processes leading to completion, yet in the gov.uk data these verbs are couched in structures that refocus them as indefinitely-lasting states rather than achievable goals.

This shift from action to state introduces a fuzziness to the meaning of these verbs. Vague language goes hand in hand with vague intentions. If there are no actions successfully completed, and no results to report, accomplishments and resultative constructions simply cannot be employed in the text. When attempts are made to say that something is actually happening when it is really only getting underway, the syntax can scarcely avoid becoming convoluted. Words such as *commit* and *work* with* find themselves incorporated into formulaic phraseological concatenations which are unlike those typically found in the language spoken by the general public. So the meaning, which is unclear in the first place, comes over as unclear as a consequence of the phraseology. The readership is unsatisfied

because its expectations of what the text ought to deliver are violated; but these expectations cannot be met because of the clash between what these documents are assumed to be for, and what is genuinely within their means to communicate.

It seems, then, that it is impossible to resolve the problem of Civil Service "gobbledygook" by using "plain language", because the point is not that the words are unfamiliar but rather that familiar words are being used to express unfamiliar meanings. "Different discourses construct different realities", Teubert (2010: 123) reminds us, and he continues: "None of these 'realities' can claim to be the only one. They do represent different ideologies. But they never represent reality as it 'really' is" (Teubert 2010: 123). In the online policy documents studied, even such plain language expressions as *working with* mean something different than they do "out in the real world", not because there is any intention to deceive, but because what *working with* means in the implementation of government policy is qualitatively different from, say, working with colleagues to achieve a mutually-beneficial goal or working with materials to build or make something concrete. When government is working with interested parties, it is collaborating with them in some way, but whereas the interest groups' goals are to make changes for the benefit of people who are directly or indirectly affected by the issue at hand, the government's primary goal can be said to be to appease those groups and thus gain (or at least maintain) consensus and future votes. *Working* suggests action towards a goal, but for government it is more important to show that it supports the goals than to bring about the desired results.

The role that metaphor plays in this "game" is subtly persuasive. In policy documents, metaphor does not perform the same kinds of rhetorical function that it does in political speeches, i.e. as a device that "can stir the emotions or bridge the gap between logical and emotional (rational and irrational) forms of persuasion" (Mio 1997: 121). Instead, it "can simplify and make understandable political events" and "resonate to underlying symbolic representations in its recipients" (Mio 1997: 121). It brings into the discourse elements which implicitly cast the government in a positive light, as active promoter of change for good, as being "on the public's side". The metaphorical lexis included in the banned words list offers accessible meanings which the public can grasp, and therefore ensures that the policy documents convey the idea that the government is indeed doing things, even when the doing has yet to materialise as actions achieved. To insist that such words should be done away with, as the Style Guide does, is not only unrealistic; it is unhelpful too. They are crucial in ensuring that otherwise "empty, meaningless text" actually conveys something to the readership it is intended to inform.

A study such as this one, which analyses a specific text type in a constrained time-frame, could be expanded by extending the time-frame and/or including a greater array of policy-related documents. Doing so would doubtless increase the

pool of data within which to search for the "banned words" and probably offer richer evidence of how they are used in this restricted language. At the same time, it is difficult to imagine that the relative frequencies of these words and the meanings that they are found to convey would change significantly. A larger data set might additionally feature some of the "banned words" that were not found in the 2013 data – provided, of course, that they are indeed used, since there is no evidence that the Style Guide was actually based on the kinds of online content it purportedly refers to. It might therefore be useful to engage in dialogue with the authors, to try to uncover what motivated their selection of banned words and to try to elucidate where and when (or indeed, if) they were encountered.

> Whenever we are unhappy with the way someone uses a word or a phrase, or with a longer piece of text, we will open a discussion about its meaning. … We may not be able to convince our interlocutor of our view. But by talking about the word … we will jointly come up with a new interpretation of it that will be added to its meaning… (Teubert 2010: 8)

Language policy is something that linguists can fruitfully contribute to. One such example is a policy paper prepared for the British Council, advising on how to use metaphor in communication with international students (Littlemore et al. 2012). Explaining how to present relatively specialised content in a way that is clear and accessible to the general public is not impossible, but the focus needs to shift away from the linguistic conservatism which has characterised style guides for civil servants over the past century and which has failed to deliver the desired results.

Acknowledgements

Linguistic projects sometimes begin in unexpected ways. The original idea came in one of many inspiring chats with my father, who passed on a newspaper clipping about the Style Guide with the words, "I thought you might have something to say about this, being a linguist". Indeed, I might.

Earlier versions of the data analysis reported in this chapter were presented at two conferences in 2014: ICAME 36 in Nottingham, and Researching and Applying Metaphor (RaAM 9) in Cagliari, Italy. On both occasions I benefitted from exchanges of ideas with colleagues: in particular Bill Louw, who made insightful comments on the relation between colligation, syntax and subtext; and John Barnden who encouraged me to look more closely at the relationship between metaphor and vague meaning.

References

Aarts, B., Chalker, S. & Weiner, E. 2014. *The Oxford Dictionary of English Grammar*. Oxford: OUP.

Bache, C. 1995. *The Study of Aspect, Tense and Action: Towards a Theory of the Semantics of Grammatical Categories*. Frankfurt: Peter Lang.

Black, M. 1962. *Models and Metaphors*. Ithaca NY: Cornell University Press.

Bougher, L. D. 2012. The case for metaphor in political reasoning and cognition. *Political Psychology* 33(1): 145–163. https://doi.org/10.1111/j.1467-9221.2011.00865.x

Charteris Black, J. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9780230000612

Charteris-Black, J. 2011. *Politicians and Rhetoric. The Persuasive Power of Metaphor*, 2nd ed. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9780230319899

Chilton, P. & Ilyin, M. 1993. Metaphor in political discourse: The case of the common European house. *Discourse and Society* 4(1): 7–31. https://doi.org/10.1177/0957926593004001002

Chilton, P. 2004. *Analysing Political Discourse*. London: Routledge.

Council of Europe. n.d. *How to Write Clearly*. <http://ec.europa.eu/translation/writing/clear_writing/how_to_write_clearly_en.pdf> (25 April 2015).

Crow, P. 1988. Plain English: What counts besides readability? *The Journal of Business Communication* 25(1): 87–95. https://doi.org/10.1177/002194368802500106

Deignan, A. 2005. *Metaphor and Corpus Linguistics* [Converging Evidence in Language and Communication Research 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/celcr.6

Fairclough, N. 2000. Discourse, social theory, and social research: The discourse of welfare reform. *Journal of Sociolinguistics* 4(2): 163–195. https://doi.org/10.1111/1467-9481.00110

Flesch, R. F. 1946. *The Art of Plain Talk*. New York: Harper.

Fischer, J. A. 2007. Why George Orwell's ideas about language still matter for lawyers. *Montana Law Review* 68: 129–149.

Flowerdew, J. 2004. The discursive construction of a world-class city. *Discourse and Society* 15(5): 579–605. https://doi.org/10.1177/0957926504045033

Fowler, H. W. 1926. *A Dictionary of Modern English Usage*. Oxford: Clarendon Press.

Fowler, H. W. 1965. *A Dictionary of Modern English Usage*, 2nd ed., revised by E. Gowers. Oxford: OUP.

Garner, B. 1987. *A Dictionary of Modern Legal Usage*. Oxford: OUP.

Gibbs, R. W. 2015. The allegorical character of political metaphors in discourse. *Metaphor and the Social World* 5(2): 264–282 https://doi.org/10.1075/msw.5.2.05gib

Gowers, E. 1954. *The Complete Plain Words. Containing 'Plain Words' and 'The ABC of Plain Words' rearranged and revised*. London: Her Majesty's Stationery Office.

Gowers, E. 1986. *The Complete Plain Words*, 2nd ed., revised by S. Greenbaum & J. Whitcut. London: Her Majesty's Stationery Office.

Grice, H. P. 1975. Logic and conversation. In *Syntax and Semantics, 3: Speech Acts*, P. Cole & J. Morgan (eds), 41–53. New York NY: Academic Press.

Gulick, L. 1984. The metaphors of public administration. *Public Administration Quarterly* Fall 1984: 369–381.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge MA: The MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Hastings, A. 1998. Connecting linguistic structures and social practices: A discursive approach to social policy analysis. *Journal of Social Policy* 27(2): 191–211. https://doi.org/10.1017/S0047279498005248

Ho, V. 2016. Discourse of persuasion: A preliminary study of the use of metadiscourse in policy documents. *Text & Talk* 36(1): 1–21.

Howe, N. 1988. Metaphor in contemporary American political discourse. *Metaphor and Symbolic Activity* 3: 87–104. https://doi.org/10.1207/s15327868ms0302_2

Johnson, E. 2005. WAR in the media: Metaphors, ideology, and the formation of language policy. *Bilingual Research Journal* 29: 621–640. https://doi.org/10.1080/15235882.2005.10162855

Kimble, J. 1996. Writing for dollars, writing to please. *The Scribes Journal of Legal Writing* 1(2): 1–38.

Kimble, J. 2002. The elements of plain language. *Michigan Bar Journal* October 2002: 40–41.

Koller, V. & Davidson, P. 2008. Social exclusion as conceptual and grammatical metaphor: A cross-genre study of British policy-making. *Discourse and Society* 19: 307–331. https://doi.org/10.1177/0957926508088963

Lakoff, G. & Johnson, M. 1980. *Metaphors We Live By*. Chicago IL: Chicago University Press.

Landau, M. J. & Keefer, L. A. 2014. This is like that: Metaphors in public discourse shape attitudes. *Social and Personality Psychology Compass* 8(8): 463–473. https://doi.org/10.1111/spc3.12125

Landau, M. J., Keefer, L. A. & Rothschild, Z. K. 2014. Epistemic motives moderate the effect of metaphoric framing on attitudes. *Journal of Experimental Social Psychology* 53: 125–138. https://doi.org/10.1016/j.jesp.2014.03.009

Lau, R. R. & Schlesinger, M. 2005. Policy frames, metaphorical reasoning, and support for public policies. *Political Psychology* 26(1): 77–114. https://doi.org/10.1111/j.1467-9221.2005.00410.x

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago IL: The University of Chicago Press.

Littlemore, L., MacArthur, F., Cienki, A. & Holloway, J. 2012. How to make yourself understood by international students: The role of metaphor in academic tutorials. *ELT Research Papers 12–06*. (Available from http://englishagenda.britishcouncil.org/research-publications/research-papers/how-make-yourself-understood-international-students-role-metaphor-academic-tutorials (3 April 2018)).

Longe, V. 1985. Apects of the textual features of officialese. *International Review of Applied Linguistics in Language Teaching* 23(45): 301–313.

Longe, V. 1999. The linguistic realization of paralinguistic features in administrative language. *Studies in the Linguistic Sciences* 29(1): 113–127.

Mio, J. S. 1997. Metaphor and politics. *Metaphor and Symbol* 12(2): 113–133. https://doi.org/10.1207/s15327868ms1202_2

Musolff, A. 1998. Metaphors and trains of thought: Spotting journey metaphors in British and German political discourse. In *Language, Politics and Society*, S. Wright, L. Hanrais & J. Howorth (eds), 100–109. Clevedon: Multilingual Matters.

Musolff, A. 2004. *Metaphor and Political Discourse. Analogical Reasoning in Debates about Europe*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9780230504516

Ogden, C. T. & Richards, I. A. 1936. *The Meaning of Meaning*. 4th ed. London: Kegan Paul.

Orwell, G. 1946. Politics and the English language. *Horizon*, April 1946: 252–265.

Partridge, E. 1948. *Usage and Abusage*. London: Hamish Hamilton.

Partington, A. 2003. *The Linguistics of Political Argument: The Spin-doctor and the Wolf-pack at the White House*. London: Routledge. https://doi.org/10.4324/9780203218259

Philip, G. 2011. *Colouring Meaning. Collocation and Connotation in Figurative Language* [Studies in Corpus Linguistics 43]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.43
Plain English Campaign. 2009. *How to Write in Plain English*. <http://www.plainenglish.co.uk/files/howto.pdf> (25 April 2015).
Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22(1): 1–39. https://doi.org/10.1080/10926480709336752
Radden, G. & Dirven, R. 2007. *Cognitive English Grammar*. Amsterdam: John Benjamins. https://doi.org/10.1075/clip.2
Richards, I. A. 1936. *The Philosophy of Rhetoric*. Oxford: OUP.
Rothstein, S. 2004. *Structuring Events. A Study in the Semantics of Aspect*. Oxford: Blackwell. https://doi.org/10.1002/9780470759127
Rundell, M. (ed.). 2002. *Macmillan English Dictionary for Advanced Learners (MEDAL)*. London: Macmillan Education.
Saarinen, T. 2008. Persuasive presuppositions in OECD and EU higher education policy documents. *Discourse Studies* 10(3): 341–359. https://doi.org/10.1177/1461445608089915
Schiess, W. 2003. What plain English really is. *The Scribes Journal of Legal Writing* 9: 43–75.
Scott, M. 2008. *WordSmith Tools, Version 5.0*. Liverpool: Lexical Analysis Software.
Semino, E. 2008. *Metaphor in Discourse*. Cambridge: CUP.
Sinclair, J. M. (ed.). 2001. *Collins COBUILD on CD-ROM v1.0*. Glasgow: HarperCollins.
Sinclair, J. M. 1996. The empty lexicon. *International Journal of Corpus Linguistics* 1: 99–119. https://doi.org/10.1075/ijcl.1.1.07sin
Steen, G., Dorst, A., Herrmann, B., Kaal, A., Krennmayr, T. & Pasma, T. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU* [Converging Evidence in Language and Communication Research 14] Amsterdam: John Benjamins. https://doi.org/10.1075/celcr.14
Steen, G., Reijnierse, G. & Burgers, C. 2013. When do natural language metaphors influence reasoning? A follow-up study to Thibodeau and Boroditsky (2013). *PLoS ONE* 9(12): e113536. https://doi.org/10.1371/journal.pone.0113536
Steinert, H. 2003. The indispensable metaphor of war: On populist politics and the contradictions of the state's monopoly of force. *Theoretical Criminology* (3): 265–291. https://doi.org/10.1177/13624806030073002
Teubert, W. 2010. *Meaning, Discourse and Society*. Cambridge: Cambridge University Press.
Thibodeau, P. & Boroditsky, L. 2013. Natural language metaphors covertly influence reasoning. *PloS ONE* 8: e52961. https://doi.org/10.1371/journal.pone.0052961
Thornborrow, J. 1993. Metaphors of security: A comparison of representation in defense discourse in post-Cold-War France and Britain. *Discourse and Society* 4(1): 99–119. https://doi.org/10.1177/0957926593004001006
Wright, O. 2013. "Only pizzas are delivered": Public sector jargon banned in first style guide for Government announcements. *The Independent*, 24 July 2013.

The individual and the group from a corpus perspective

Michael Barlow
University of Auckland

The connection between the individual and the group has been a recurring issue since the earliest investigations into the nature of language. The contribution made here is to provide some empirical data on individual differences in the production of phrases and constructions. For this purpose, we examine some aspects of the speech of five White House press secretaries. The results show that individual speakers reuse favoured constructions and routines, leading to a clear disparity in the frequency of common lexical and grammatical patterns among the different press secretaries. Investigating the extent of idiolectal variation provides a baseline for assessing corpus results based on aggregate data and for understanding the connection between the individual and the group.

1. Introduction

> Our individuality only exists in exercising our otherness inside a discourse community. (Teubert 2010: 136)

From the beginning of modern linguistics a distinction has been made between the language of an individual – both the tangible language output and the associated underlying cognitive system – and the language associated with a linguistic community, which is typically viewed as a generalised abstract description covering the conventional linguistic patterns (or signs) belonging to a group of speakers. One early and clearly very influential distinction along these lines is Saussure's *parole* and *langue*, which can be seen as setting out the foundations for structural linguistics and developing the idea of language as an abstract system associated with a social group. The seemingly contradictory position that grammar must be linked both with the individual and with the group has to be faced, implicitly or explicitly, in all linguistic theories and practices. Different linguistic paradigms have tried to ameliorate the problem by focusing on particular aspects of behavior or knowledge related to language use. Theoretical linguists following

https://doi.org/10.1075/scl.87.07bar

the generative paradigm take a mental perspective on the study of grammar and have tended not to be too concerned with *parole* or *langue* but assume an idealised social representation of language, which is conveniently located in the head of every speaker of the language. Other sub-disciplines, such as sociolinguistics, come face-to-face with the speech of individuals in their linguistic investigations and the tension between *parole* and *langue* is more immediate.

In this chapter I take some steps in investigating the relation of the individual to the group in general terms through a corpus analysis. In discussions of the relation of individuals and communities, there are two main perspectives: the assumption that the language associated with a discourse community is an amalgamation of the individual contributions or that the community grammar has a privileged status and the grammars associated with individuals are some version, perhaps aberrant, of this community grammar. In both cases, grammar is viewed abstractly as a single system covering comprehension and production. What has been missing to a large extent is empirical data on the difference between individual productions – spoken output – and the general grammatical forms of a discourse community. The distinction between production and comprehension is difficult to investigate, but as a factor in the relation between the individual and the group, it should not be overlooked. To take some initial steps in probing this distinction, I use a corpus to examine individual patterns in language use – what might be called 'idiolectal grammar'. Establishing the nature and extent of degree of individual variation provides a baseline for assessing corpus results based on aggregate data and for understanding the connection between the individual and the group.

In this chapter I examine the variation in individual contributions to the discourse and I present some representative data below, I approach this topic rather broadly: placing most emphasis on frequent grammatical patterns rather than search for the most distinctive words or phrases used by particular individuals. Such expressions might be of interest due to their particularity, but they represent features at the periphery of grammar that are clearly idiosyncratic. While idiosyncrasies are potentially of interest and do in some sense define idiolectal differences, the aim here is to investigate what can be considered to be central patterns of grammar and show in an initial way the extent to which grammars vary from speaker to speaker. This focus on idiolectal grammar is introduced to add to and contrast with the many types of investigations of the language associated with different speech communities.

I explore below some of the perspectives taken on the grammar of the individual and the group. Many discussions on this issue involve various kinds of abstractions and idealisations. While abstractions may be useful and to some extent necessary, the use of empirical data provides specific information on the usage patterns of individuals and groups.

2. The individual and the group

Sapir (1973: 357) lays out two approaches to the relation between individuals and groups:

> To some [psychologists and sociologists] the group is a primary concept in the study of human behaviour; many sociologists say that the individual has no reality, aside from his biologically defined body, except as a carrier or crystallizer of meanings that are derivative of group action and interaction. To others, however, the individual remains as the sociologically primary entity and groups are more or less artificial constructs which result when individuals, viewed as essentially complete physical and psychological entities, come into contact with each other.

Sapir is describing different philosophical positions underlying approaches to the individual and the group within psychology and sociology. The same opposing positions are evident when examining the language of the individual and the group. In discussions on this topic a variety of issues arise, which while following a similar conceptual path are somewhat distinct. Among the different strands in the idiolect-sociolect debate are issues relating to conventionality and communication, the primacy of idiolects/sociolects, and variation in idiolects/sociolects.

According to Saussure, the representation of language in the individual comes from the accumulation of "impressions that are perceptibly the same for all" (Saussure 1985: 33). The connection between the individual and the social is described as follows:

> If we could embrace the sum of word images stored in the minds of all individuals, we could identify the social bond that constitutes Language (Langue). It is a storehouse filled by the members of a given community through their active use of speaking, a grammatical system that has a potential existence in each brain, or more specifically, in the brains of a group of individuals. For Language is not complete in any speaker; it exists perfectly only within a collectivity. In separating language from speaking we are at the same time separating: (1) what is social from what is individual; and (2) what is essential from what is accessory and more or less accidental. (Saussure 1985: 33)

This quote captures several recurrent ideas and dilemmas within linguistics. There is the idea that each individual in the community is exposed to essentially the same language and this leads to individual grammars, which in key aspects are identical. In addition, there is the attractive notion that if we can aggregate the language of individuals to an appropriate degree of abstraction, then we arrive at a community grammar. The linguist as analyst is here taking a scientific position that parallels in some sense the behaviour that is observed. We know that each individual has a representation of grammar and we observe individuals interacting successfully

with other speakers and so there must be some commonality (expressed to different degrees) that links all the speakers of the language.

Reflecting a widely-held view, another idea expressed in the quote is that it is only in the collective or community that the true or perfect grammar is represented. In this vein, Labov (1989) has consistently argued that the language of the group is the proper object of study. He states "language is not a property of the individual, but of the community. Any description of a language must take the speech community as its object if it is to do justice to elegance and regularity of linguistic structure" (Labov 1989: 52). The first statement can be taken to be an assumption or axiom related to the sociolinguistic tradition instigated by Labov. While the second statement can be interpreted in a variety of ways, it is clear that the main idea is that regularity is associated only with the language of the group and it might be the case that individuals, as members of different intersecting groups, will exhibit essentially random irregularities. However, we can note that the statement relates to those aspects of language that Labov finds interesting from a sociolinguistic viewpoint.

Teubert (2010: 128) equates Saussure's *parole* with discourse, "the entirety of all the extant and accessible texts (verbal utterances) entered by the members of a specific discourse community." What is real for Teubert is not the representation of language in the mind nor some abstract *langue* but the combination of utterances (and texts) within a context, which can be viewed as "distributed collective mind" (Teubert 2010: 135).

The notion of an individual grammar should naturally align with the cognitive perspective – the knowledge of language possessed by each language user – since cognition is essentially a property of individuals. While we can accept that language use has an important social dimension, it is evident that social interactions lead individuals to possess knowledge structures or processes concerning the structure and function of language, including how social context comes into play in language use. Cognitive development crucially takes place within a social context, with various joint or cooperative actions including language interactions. The latter perspective emphasises the social side of individuals: their knowledge of and their interaction with others in a variety of styles and settings. The notion of socialised cognition is one view of the social in language. We can refer to the 'sociocognitive individual', a concept that highlights the cognitive structures typically associated with grammar and social knowledge, arising out of spoken discourse and interactions with other members of a speech community.

To avoid confronting idiosyncratic behaviour, many linguists have taken an idealised grammar associated in an unspecified way with a speech community and projected that grammar onto individual speakers. Thus, even if the focus is on the cognitive aspects of language, there is a notion that the proper representation of

language lies beyond the individual and this leads to the idealised speaker/hearer, representing the system found with the collective, as noted by Chomsky in his earlier writings (Chomsky 1965: 3).

As is well-known, Chomsky's influence dramatically shifted the theoretical paradigm and hence the practice and focus of research in linguistics. One of the changes in direction came from his argument that linguistics is a branch of cognitive science and that there is a specialised language faculty alongside other cognitive systems, such as face recognition. Different versions of generative grammar have all maintained this view of language as a cognitive system with a similar disinterest in the social dimension of grammar. In investigating the core features of language within a cognitive domain, it is the social aspect that is viewed as peripheral and idiosyncratic.

A slightly different tack is taken in the theoretical framework of Head-driven Phrase Structure Grammar (HPSG) in that Saussurean signs are incorporated directly into the formalism. Pollard and Sag (1994: 15) explicitly introduce the concept of signs and render them within the theory using feature structures consisting of "phonological, syntactic, semantic, discourse and phrase-structural information." Thus some of the social, conventional aspects of language materialise in the representation of grammatical knowledge in the individual.

Alternative views of language as a cognitive system can be found in cognitive linguistics (Langacker 1987) and similar research paradigms (Goldberg 2006, Lamb 1999, Talmy 2000). Rather than view grammar as a system with its own very distinctive properties, as generativists do, language knowledge is seen as having the same underpinnings as other cognitive processes in the brain. It should be noted that recent developments in this strand of linguistics have shown more interest in social factors (Croft 2009).

Another perspective is captured by the phrase "the language of the group". We recognise that there are dialects and languages, which are loosely tied to regions, large or small, or to other social groupings. And there is some sense that there is a grammar shared to a large extent by all members of the relevant language group. On a smaller scale it is clear that people have multiple interactions and social ties within a small network of people. Beyond this, however, it has proved to be very difficult to pin down what is meant by a speech community (or dialect or sociolect) and it is evident that research in sociolinguistics encompasses a range of notions of 'community', which include anything from a small close-knit village to a community of practice to a regional dialect.[1]

1. See Patrick (2002) for a discussion of the different ways that speech community has been defined within the sociolinguistics studies.

There are compelling reasons for seeing language as a social behaviour. Languages are acquired in a social setting and it is at the level of the group that conventions are negotiated and communication takes place. Language is not primarily private (Jakobson 1971); it is mainly used for communication within a group. Based on these facts, the argument can be made that language must be a property associated with the group and not with the individual. In other words, we can say that the key aspects of language of interest to linguists come from the group and not from the individual. Therefore, there must be some language, some abstraction, or some norms, associated with the group. As mentioned above, one well-established argument for a social view of language as the appropriate object of study for linguists goes back to Saussure and rests on the idea of meaning as a set of conventions supported by a social group. Labov (1989: 52) states that "individual behaviour can be understood only as a reflection of the grammar of the speech community." The social group is the only level where regularities emerge and that if idiolects are examined, the analyst is confronted only with idiosyncrasies and the coherence associated with the language of groups is obscured.

Thus, some perspectives may make us wonder whether it is worthwhile examining variation in the individual. Some variation within individual speech will be linguistically uninteresting because it reflects changes in expression resulting from the particular set of communicative situations that the individual takes part in. Other aspects of variation, however, may, by their nature, only be discernible at the group level. Nevertheless, a knowledge of idiolects can provide a sense of the background variation against which variationist and discourse studies that examine group variation and discourse-driven patterning can be assessed.

As noted above, discussion of the relation of language of individuals and groups have been based on a view of grammar as a single system and have not distinguished production and comprehension. We might not want to envisage a separate production grammar and a comprehension grammar for individuals, but that would be a useful corrective to a simple focus on "grammar". For now, we can surmise that the comprehension grammar of individuals is close to the grammar of the collective. The nature of individual differences in production is something that we can study emprically.

2.1 Research on individual speech

Most sociolinguistic studies quite naturally focus on sociolects, but there are also a variety of investigations of individual speech. Dorian (1994: 562) analysed the language of members of a small isolated fishing village with a close-knit community and found that "individuals who show similar patterns in the case of one variable do not necessarily show similar patterns in the case of another." She notes that

"this individuality in the patterns of varied usage is a defining feature of personal pattern variation" (Dorian 1994: 562). Dorian (1994: 686) goes on to voice the suspicion that this kind of individual variation is not generally detected because of the data collection methods in dialectology and variationist sociolinguistics. Johnstone and Bean (1997: 235) make a similar point and state that "theories of language that begin with groups, rather than with individuals … make it difficult to hear particular voices, and harder to take them seriously as part of the explanation of variation." Fischer (1964: 486) refers to Gauchat (1905) as a classic study of individual differences. Gauchat investigated phonological differences in a small French-speaking area in Switzerland. Fischer goes on to comment on the relation between the group and individual variation and takes the position that the idiolect is primary:

> What I am proposing might be called comparative idiolectology rather than dialectology. Ideally, a thorough description of a single dialect would be based on the study of a sizable sample of the idiolects in a local speech community, in the same way that a thorough description of a language would be based on the study of a sizable sample of its dialects. (Fischer 1964: 487)

Milroy (1987: 131) observed differences in phonological variables known to be linked to age, sex and area for two middle-aged women who shared many characteristics and had very similar backgrounds. This led Milroy to explore the extent to which these differences could be explained in terms of differences in the strength and density of social networks associated with the different individuals. She found that generally there was a correlation between individual scores for the linguistic variables and networks scores.

Johnstone (1999, 2000) has carried out a considerable amount of research on the language of individuals, but these studies have tended to focus on the characteristics of individual speakers rather than their grammars. The focus is on the individual and self-expression: the way in which the individuals use linguistic resources to present themselves and position themselves within a discourse. This research considers the resources available to the individual and the way in which individuals make choices. The result of these choices could be said to culminate in an individual style, or viewed from a different perspective, can be seen as developing different voices for different situations. Approaches that see speaking as an "act of identity" (Le Page & Tabouret-Keller 1985) clearly put the speaker at centre stage. Thus rather than see an individual as being bound by sociolinguistic variables, the view is that individuals use variability to actively position themselves within a social setting: revealing "their personal identity and their search for social roles" (Le Page & Tabouret-Keller 1985: 14).

2.2 Individuals and groups in corpus linguistics

The approaches referred to in the previous sections are either based on intuition or on small data samples and hence the move to corpus data must be a step in the right direction for an investigation of the language of groups in the sense that a corpus typically contains a large sample of the language produced by many people and that the language sample is designed to be representative of some aspect of usage. In addition, individual contributions within the corpus are potentially accessible. Nevertheless, problems remain and the use of corpus data is not automatically going to provide insights into the language of individuals and groups and the relation between the two.

Some corpora, more than others, are designed to reflect the language of particular groups and there are projects such as the *International Corpus of English* (ICE), which collects corpora representing different Englishes. The collection currently covers around 14 varieties of English, from Ireland to India, and facilitates the comparison of the different varieties. The language contained in the corpora can be said to represent the language that Irish speakers of English or Indian speakers of English might be exposed to or would recognise rather than necessarily the language that they produce.

Elsness (2013) used a variety of corpora, including some from the ICE, to investigate some characteristics of the different varieties, including the proportion of use of *he* versus *she*, the verbs used following the pronoun, and various cultural terms in different varieties of English. He discusses the differences found among the different corpora and poses the question faced in this kind of corpus study:

> A further intriguing question which remains unanswered is to what extent the recorded figures reflect real differences in and between the societies depicted, or merely differences in what fragments of each culture are reported in the kinds of text on which the investigation was based. (Elsness 2013: 134)

The *British National Corpus* (BNC) was not designed in such a way as to be an accurate sample of language use in Britain in the 1990s and the comment by Elsness is also something to consider for a large, general corpus like the BNC, which is around 90% written texts and 10% spoken texts. One advantage of the BNC is that it contains some metadata on the component files that can be used to investigate sociolinguistic questions based on the major variables of age, gender, region, and socioeconomic status. Baker (2010) describes some sociolinguistic studies based on the BNC but also points out some problems concerning sparse data, representativeness and inter-speaker variation. When focussing on gender in a corpus, the researcher is typically gathering a sample of men and women without much information concerning other variables, including not only age, region and

socioeconomic group but also more generally the setting. This means that we cannot be certain whether differences found in a corpus are actually related to gender or whether the populations of males and of females differ, perhaps more significantly, in some other feature such as age.

As noted above, the advantage of using a corpus comes from the ability to get a reasonably-sized sample. A disadvantage is that corpora tend to be context-poor and we are unlikely to know the setting or situation in which the men and women are speaking/writing. Baker points out that it may be a difference in the situation or setting rather than the difference in gender that leads to a distinction in language use.

Following up on the issues raised by Baker noted above, Brezina and Meyerhoff (2014: 7) review corpus-based sociolinguistic studies in order to determine "to what extent the findings represent a reliable reflection of the social reality and to what extent they suffer from the shortcomings of the aggregate data methodology." They use various statistical analyses that demonstrate clearly the need to take account of individual speaker differences as well as social groupings.

Thus there are two major problems facing corpus studies of the group. One is that the corpora do not readily coincide with a group. It is not going to be the case that the contributors to the BNC form any kind of meaningful linguistic group. The second problem is that aggregate data of the sort typically derived from corpus analysis, the average or normalised frequency for use of a particular word or construction might be an approximation of the language of some group – British speakers – but it is not, or is very unlikely, to represent the usage of individuals. It may not even correspond to the most typical usage. And if we make use of the metadata on individuals, we cannot be sure of the influence of the different characteristics or variables, and again, we cannot access information on the discourse communities within which the individuals are interacting.

The *Corpus of London Teenage Language* (COLT) might be one example of a corpus that is more closely associated than most with a specific community. However, the recordings come from five different boroughs of London and represent different socio-economic groups. And, of course, the teenagers talked to a variety of adults so while the corpus can be used to extract some features of the language of London teenagers, that language does not correspond to a particular social network and it includes, necessarily, examples of the speech of individuals outside the teenage groups.

The language of individuals may be retrievable from a corpus. In spoken usage, individual speakers may be annotated with a distinct tag and in writing each file may correspond to the output of one writer. Whether theoretically the individual contributions could be identified is a moot point, however, there is typically little interest in retrieving this kind of information. There are some exceptions, of

course. An article written by Coniam (2004) is titled *Concordancing Oneself* and is based on an analysis of the author's own writings. In addition, Mollin (2009) collected a 3 million-word corpus of the output of Tony Blair and used that to provide insights into the features of his idiolect, especially with regard to collocations that could be counted as "Blairisms".

What is missing from the discussion is a consideration of the implications of the bidirectionality of influence between individuals and the group. Teubert (2010: 136) notes: "One way to make sense of the discourse is to look at it as something emerging from the collaboration of individual minds. These minds in turn are continually re-creating themselves through the symbolic input they receive." Language input from a discourse feeds or creates the language knowledge of an individual. And the individual's contributions add to the language of the group. However, the crucial consideration is that the input from the group is associated with what we can call the receptive or comprehension component of the grammatical system. While the output from the individual is based on the grammatical routines linked to production.

Since corpora are amalgamations of the speaking or writing of different people, the patterns that are extracted from corpus data tell us primarily about language that individuals are exposed to and hence lead to grammatical descriptions that correspond to a model of comprehension. It is generally assumed that productive patterns are some subset of comprehension patterns of language, but this notion is very vague and only with a detailed knowledge of idiolectal data is it possible to understand the connection between comprehension and production in grammatical terms. It is well-known that language users comprehend a wider range of linguistic expressions than they produce. There is, however, an assumption that there is essentially one grammar covering comprehension and production and while we might comprehend but not produce specific instances of jargon or instances of Shakespearean English, it is assumed that generally there is a considerable overlap between comprehension and production: most of the things that we hear, we also say. I suggest that this is not the case. We have our own favoured routines for specific purposes in specific situations and we tend to stick to those, over a period of several years at least. These routines serve their purpose. We use them to express ourselves and convey our thoughts. We are aware that others employ different routines. We expect that and may occasionally "borrow" a useful piece of discourse.

It may well turn out that there is considerable variation in the extent and frequency of borrowing. Some people, like myself, might be quite set in their ways and only slowly make changes, perhaps saying *hey* instead of *hi* under the influence of undergraduate students at an American university. Others may much more easily take snippets of discourse that they hear and reproduce them as their own.

We see in the data presented below an indication that individual speakers have their preferred routines for conveying their ideas. Presumably, these preferences with respect to production have little or no effect on the speakers' ability to interpret the meaning of their interlocutors. Let us take a simple example. The utterances in (1) and (2) come, respectively, from a graduate seminar in the humanities[2] and from a White House press conference.

(1) Having said that and assuming that I'm right, does Cort still have some kind of point?

(2) Having said that, the president doesn't want Guantanamo open any longer than it has to be.

The device of using *having said that* to introduce a contrast is not particularly unusual and would appear to be quite normal in any sort of commentary within a professional setting. It is not difficult to imagine that we use this phrase from time to time since it is so ordinary and its function is transparent. The fact that this is a perfectly normal phrase notwithstanding, I claim that this is an example of a phrase that some speakers use in production in an appropriate professional setting and some do not.

If we examine the speech of five previous White House press secretaries in press conferences, we find evidence that three of them use *having said that* as part of their repertoire. Two of the press secretaries have just one instance and the case of one of them, we have quite a large sample of 600,000 words of their speech, yielding this one instance. And looking at the press conference transcript we find that one of the media representatives also uses the phrase and so there may be some kind of priming or temporary use on the part of this speaker. It is reasonable to assume the function of this particular phrase to express a contrast is accomplished using some other linguistic device or devices. For two of the press secretaries there is no instance at all of use of the phrase and for one of them we have 1.2 million words of their spoken output covering a couple of years of press conferences. Given the nature of the daily interactions, this press secretary must have heard and processed the phrase numerous times. And yet, he does not himself use it, or uses it only rarely.

To pursue the empirical part of the study, we examine in more detail the speech of the five White House press secretaries. There are several advantages in using this kind of data. One is the fact that the transcripts, which are sufficiently accurate representations of the spoken interactions for the purposes of the present study, are available and readily accessible. Second, the amount of speech transcribed is

2. The example comes from the MICASE corpus.

considerable and here we will work with individual speech corpora that range between 200,000 and 1,200,000 words of running text. A third and very important advantage of working with this dataset is that the context of the discourse is held constant across the different samples. The content changes, of course, but the overall format of press conferences does not change and virtually all the discourse involves the press secretary being questioned closely by members of the media. One disadvantage of the single setting used here is that since we do not have access to the speech of the White House press secretaries in more informal situations, and so we are only able to obtain a partial picture of idiolectal usage.

The sources of variation that contribute to the speech of individuals are wide-ranging. The content of the discourse, type of discourse interaction, the situation, the characteristics of the interlocutors, processing constraints, etc. all have a role to play. It is important to lessen the influence of these factors by investigating the speech of speakers within the same setting, press conferences, in this case, and by using fairly large sample sizes to minimise the effects due to changing content, etc.

The five press secretaries chosen for the study are: Mike McCurry (1994–98), Ari Fleischer (2001–03), Scott McClellan (2003–06), Tony Snow (2006–07), and Dana Perino (2007–08). The data presented here extends the results reported in Barlow (2013), which contains more information on the data and methodology. The transcripts were obtained using PHP scripts which automatically accessed and downloaded the relevant web pages and then extracted the transcripts. In order to reduce the possibility of priming from one day to the next, transcripts on consecutive days were not chosen, and, in fact, the typical interval might be two or three days and occasionally longer. The files were tagged for POS using the CLAWS7 tagset at the Wmatrix site[3] (Rayson 2008). The selection of these particular press secretaries was made mainly on the basis of the length of their tenure since one objective of this study is to work with reasonably large samples of individual usage. The sampled data covers around a year or more of speech output for each speaker. The press conferences occasionally contain a short statement made by the press secretary or some other White House official, but mostly the format involves answering questions posed by journalists. The length of each press conference varies, but they usually last 30 to 50 minutes. To make it straightforward to compare the spoken output of the press secretaries, the transcripts were edited in order to remove all the speech not belonging to the press secretary. Then the speech samples were split into text files containing 200,000 words each. The makeup of the collection is shown in Table 1.

3. <http://ucrel.lancs.ac.uk/wmatrix/>

Table 1. Composition of the collection

Press Secretary	Number of Samples	Total word count
Tony Snow	1	20,000
Mike McCurry	6	1,200,000
Scott McCllelan	3	60,000
Ari Fleischer	4	80,000
Dana Perino	1	20,000

To follow up on the point made above concerning the wide range of variation in the use of the very ordinary phrase, *having said that*, I examined the frequency of use of some other common phrases. Table 2 lists the five press secretaries and shows for each of them the number of instances per 100,000 words of three phrases: *having said that, on the other hand*, and *as I said*.

Table 2. Frequency of use of some lexical phrases by White House press secretaries

Press Secretary	*having said that*	*on the other hand*	*as I said*
Tony Snow	6.5	72.5	16.5
Mike McCurry	0	2.8	8.5
Scott McClelland	0	0	17.7
Ari Fleischer	3.5	0.8	13.5
Dana Perino	0.5	0.5	27.5

All the press secretaries use the phrase *as I said*. Dana Perino, however, uses it 3 times more frequently than Mike McCurry. If we look at *on the other hand*, we find that one of the press secretaries does not use it in a 600,000-word sample. Two other press secretaries use it very rarely and Mike McCurry only uses the phrase about three times per 100,000 words. This minimal usage can be contrasted with that of Snow who uses *on the other hand* about 72 times in 100,000 words.

These are somewhat random examples chosen to highlight how, even within a discourse community, we have a state of affairs in which the phrase, which appears perfectly normal for the context and is not marked in any way, is not part of the spoken repertoire of all the participants. As noted, this is merely illustrative and the general point is that the distinction between the language comprehended and the language produced by individual speakers is greater than has been assumed.

In what follows the aim is to investigate variation in individual production by assessing the frequency of use of several lexicogrammatical constructions. It is worth considering what we might expect to find when we look at the frequency of use of linguistic constructions by the press secretaries over a period of months or years. Since the content of the press conferences changes and the discourses

vary from one press conference to the next, we might expect to find a more or less random pattern of use of different constructions that follow in some manner the flow and content of discourse in each press conference. The press secretaries are for the most part responding to questions from the media and so we would expect to find some natural level of variation from one press conference to the next and whether it is the same or a different press secretary taking part would appear to be a minor consideration in terms of the language used. As we will see, however, individual speakers maintain their own preferred ways of speaking despite the factors alluded to above.

3. Results

In the following, two features will be investigated: negation and present perfect tense, which will be followed by a correspondence analysis.

3.1 Negation

The first probe we will look at is negation (*not* and *n't*). The null hypothesis is that the use of negation is mainly determined by content and interactions: the type of questions and the form of the questions posed by the press corps. The frequency of use of negation in the different samples is shown in Figure 1. We find that there are around 2000 instances of *not* or *n't* per 200,000 word sample. It turns out that in this case Snow and Perino use negation the most and McClellan's output contains the fewest negative expressions. The difference between McClellan and Snow is quite remarkable, with Snow using negation at twice the rate of McClellan. This situation in which the high frequency use by the individual who has a marked preference for the pattern is double that of the individual with the lowest frequency of use is quite common (Barlow 2013). It is difficult to conceive of contextual factors that could lead to such a disparity in use. It might be argued that press secretaries are often issuing denials of various sorts and that the frequent use of negative expressions by Snow can be explained by his being confronted with an above average number of comments leading to this high incidence of use of negative expressions. It seems unlikely though that the period of Tony Snow's reign was particularly contentious. Moreover, it is important to stress that the samples are not based on a single press conferences where variation due to the nature of the topics discussed and the manner in which they are discussed is to be expected. Each 200,000 word sample contains around fifty separate press conferences, spread over a period of about six months.

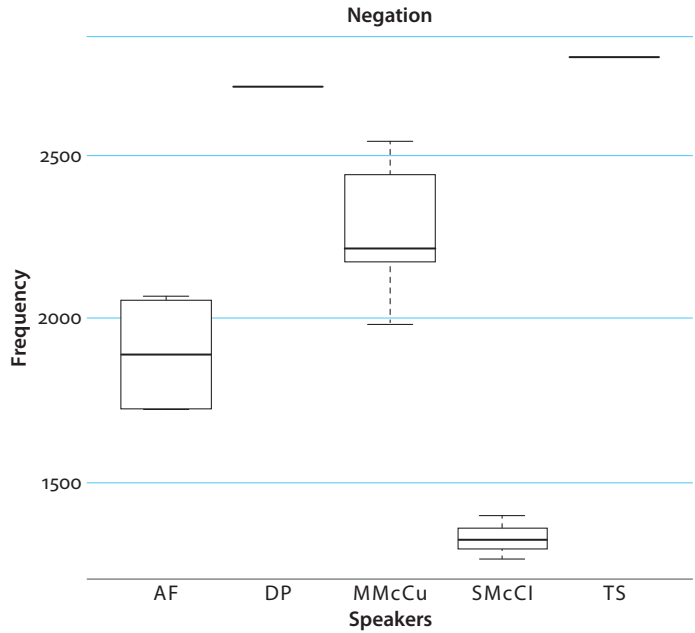

Figure 1. Frequency of use of *not/n't* by White House press secretaries
AF=Ari Fleischer, DP=Dana Perino, MMcCu=Mike McCurry, SMcCl=Scott McClelland, TS=Tony Snow

The most likely explanation is that the difference in the rate of use of negation by the two press secretaries is due to differences in their preferred grammatical routines and can be described as differences in their individual grammars related to language production. Further evidence for this comes from the remarkable stability in the frequency of use of negation by McClellan in the three separate samples, as seen in the boxplot in Figure 1.

The usage by Snow and Perino is almost identical with respect to frequency of occurrence. We only have one sample file for each of them and so we cannot discount the possibility that further data points would reveal a high degree of variation, but for the speakers for whom we have several samples, the variation is surprisingly low and we have no reason to suppose that the results for Snow and Perino would be any different. The fact that the usage is similar is not surprising. We do not expect all individuals differ with respect to all constructions. What we do expect is that Snow and Perino will differ with respect to some other parameters.

Looking only at the frequency of the use of *not/n't* can only give a broad brush picture of constructions involving negation but again it seems reasonable to suppose that if we examined the frequency of constructions involving negatives, we would see a similar picture in terms of individual differences.

Given the nature of intra-speaker and inter-speaker variation in the results displayed in Figure 1, it seems unlikely that the patterns found are due to some random sampling factor. To show that this is indeed not the case, we take six 200,000-word samples made up of the speech of multiple individuals from the press corps and examine the variation in the six samples, randomly labelled A, B, C, D, E, F. In order to make a comparison with the data from the individual speakers, a pair of samples are randomly combined and represented as a box plot, comparable with those shown in Figure 1. The results are shown in Figure 2. As can be seen, the median value of the samples is quite consistent from sample to sample and the difference between Figures 1 and 2 confirms that inter-speaker variation is quite marked. It is worth noting that the results shown in Figure 2 reflect what would be seen in a typical corpus-based study in which individual data is not identified.

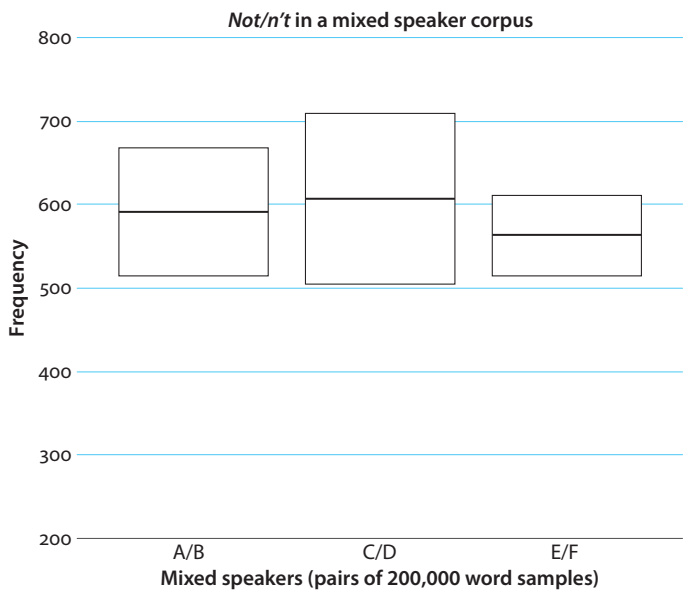

Figure 2. Frequency of *not/n't* in a corpus containing several speakers

What these results show is that despite the differences in the form of the language of their interlocutors and despite differences in the topic of the press conferences, each press secretary contrives to use negation markers and negation constructions within the setting of press conferences at a more or less consistent rate over several months or years. In the terms of Teubert (2010) we can say the reuse and replication of the discourse varies from individual to individual. One might reasonably expect to have found a much greater range of variation in use of the language expressions and a considerable degree of overlap in frequency of usage among speakers. There

is, of course, some overlap in the frequency of use of negation markers and it is not the case that each speaker occupies a completely distinct place in the frequency bands. For instance, McCurry used negation more frequently than Fleischer, but there is considerable overlap in their distribution ranges. However, a more fine-grained analysis, based on different constructions involving negation is likely to reveal the different preferences of the two speakers.

3.2 Present perfect

The next grammatical probe we examine is the frequency of use of the present perfect tense. We find in Figure 3 that McCurry, McClellan, and Snow exhibit similar, relatively high-frequency usage of this tense and there appears to be quite a broad range of variation in use of the present perfect over time.

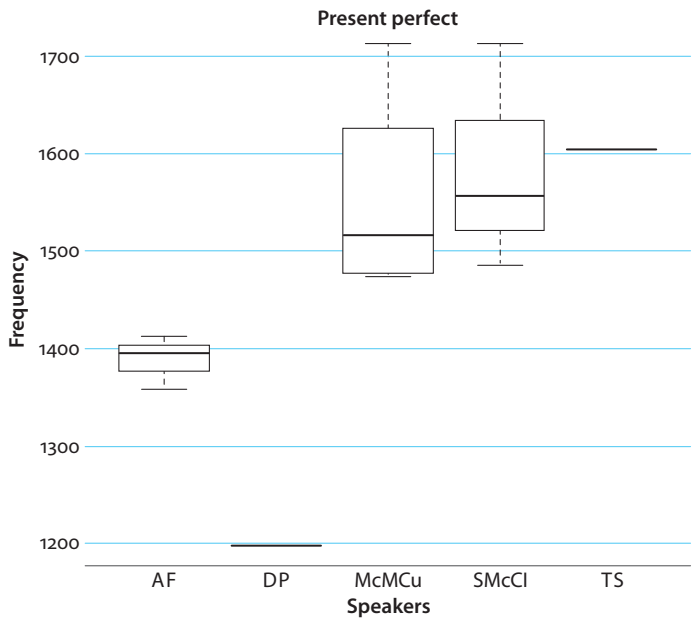

Figure 3. Frequency of use of present perfect tense by White House press secretaries AF=Ari Fleischer, DP=Dana Perino, MMcCu=Mike McCurry, SMcCl=Scott McClelland, TS=Tony Snow

Looking in more detail at the frequent verbs that occur in the present perfect tense, we discover some distinctions among the individual speakers, although also with some overlap. The results for the verbs *get* and *say* are shown in Figure 4.

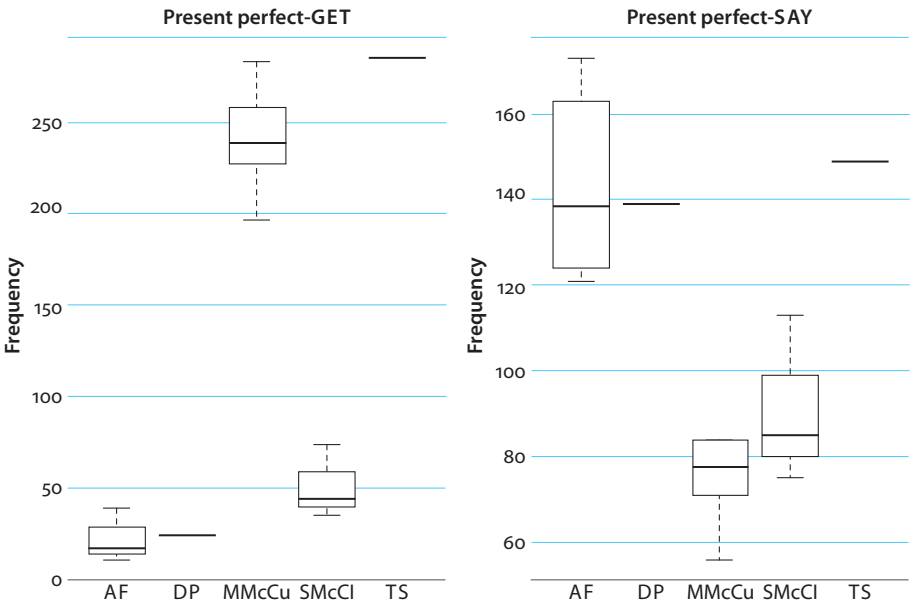

Figure 4. Frequency of use of *get* and *say* in present perfect tense
AF=Ari Fleischer, DP=Dana Perino, MMcCu=Mike McCurry, SMcCl=Scott McClelland, TS=Tony Snow

The general usage of the present perfect by McCurry and McClellan is very similar, as noted above. When we look at particular verbs, we see that for the verb *say* McClellan uses it slightly more frequently than McCurry, but they are in the same general range of use, as indicted in the right-hand graph in Figure 4. With respect to the use of the verb *get*, however, there is a clear difference between the two speakers. The box plot on the left in Figure 4 shows that McCurry uses *get* in the present perfect nearly five times more frequently than McClellan.

Again, we can contrast these results with counts of use of present perfect in a corpus made up of many speakers, as shown in Figure 5. We can see that the pattern of sample variation in the corpus made up of multiple speakers is nothing like the pattern we have seen for individual speakers.

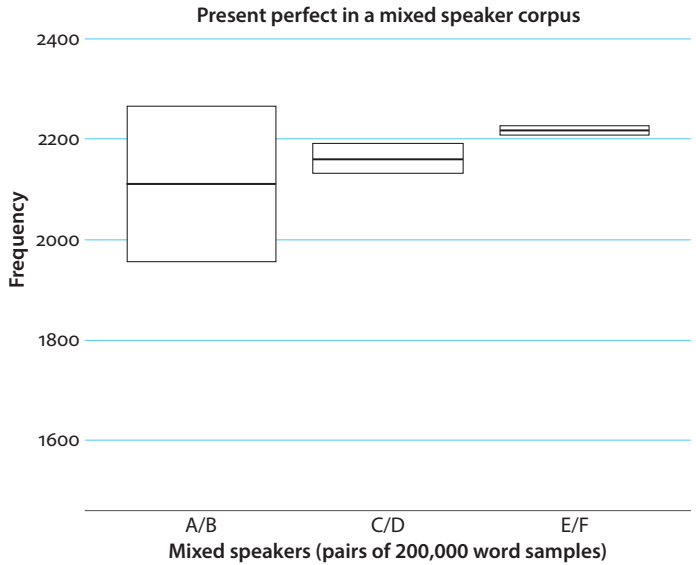

Figure 5. Frequency of present perfect in a corpus containing several speakers

3.3 Correspondence analysis

Finally, let us explore the use of a correspondence analysis to give a visual representation of individual production. If we take just four of the very general constructions – passive, present perfect, negation (*not/n't*) and '*it* BE ADJ *to*' – and use the frequency data for each of the 15 samples of speech of White House press secretaries, we can produce a 15 by 4 contingency table. This can be analysed such that the distances between the rows and the distances between the columns are calculated (according to the chi-squared distance measure). The resulting two sets of distances are superimposed and reduced to a two-dimensional map (see Baayen 2008: 129 for details). This map provides a visual representation of the similarities among the different speaker samples.

The results of this particular analysis, based on the four constructions, is given in Figure 6. The main finding of interest here is that the file samples for each speaker tend to cluster together. Thus the four sets of frequency data can quite easily distinguish the samples from Fleischer (A1-A4) from the samples from McCurry (M1-M6), for instance. The position of the constructions on the map shows the relative "pull" of the different files. For instance, we see the affinity between the passive and the samples of the speech of Fleischer.

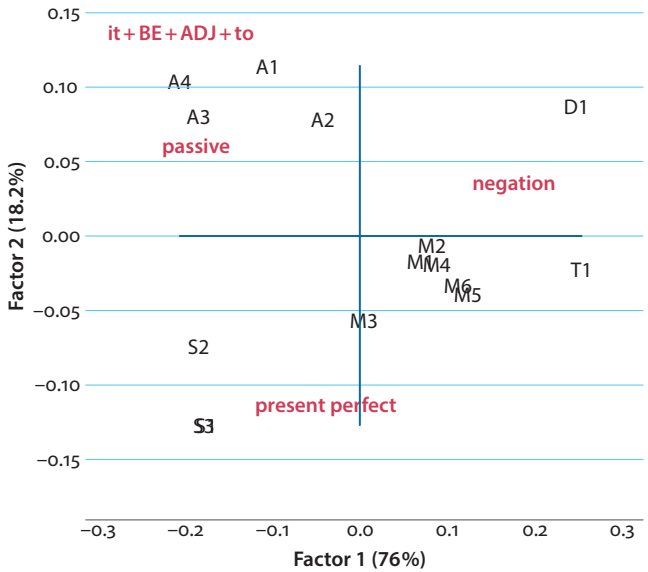

Figure 6. Grouping of sample files using a correspondence analysis

4. Conclusion

In this chapter, I have explored the tension between the language of an individual and the language of the group. While this issue has not been resolved, I have provided some empirical data giving a sense of the nature of individual differences in the frequency of usage of phrases and constructions in spoken output. These differences point to the need to separate production and comprehension when considering the broader picture concerning the individual and the group. The unity associated with a community includes a range of phrases and constructions and individual members will themselves favour particular expressions and avoid the use of others. The empirical data here is not extensive, but patterns found are consistent with the findings reported in Barlow (2013). It is clear that individual speakers reuse favourite constructions and routines, leading to a disparity in the frequency of lexical and grammatical patterns among the different press secretaries.

References

Baayen, H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511801686

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: EUP.

Barlow, M. 2013. Individual differences and usage-based grammar. *International Journal of Corpus Linguistics* 18(4): 443–478. https://doi.org/10.1075/ijcl.18.4.01bar

Brezina, V. & Meyerhoff, M. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1): 1–28. https://doi.org/10.1075/ijcl.19.1.01bre

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge MA: The MIT Press.

Coniam, D. 2004. Concordancing oneself: Constructing individual textual profiles. *International Journal of Corpus Linguistics* 9(2): 271–298. https://doi.org/10.1075/ijcl.9.2.06con

Croft, W. 2009. Toward a social cognitive linguistics. In *New Directions in Cognitive Linguistics* [Human Cognitive Processing 24], V. Evans & S. Pourcel (eds), 395–420. Amsterdam: John Benjamins. https://doi.org/10.1075/hcp.24.25cro

Dorian, N. C. 1994. Varieties of variation in a very small place: Social homogeneity, prestige norms, andlinguistic variation. *Language* 70(4): 631–696. https://doi.org/10.2307/416324

Elsness, J. 2013. Gender, culture and language: Evidence from language corpora about development of cultural differences between English-speaking countries. In *English Corpus Linguistics: Variation in Time, Space and Genre*, G. Andersen & B. Kristin (eds), 113–138. Amsterdam: Rodopi.

Fischer, J. L. 1964. Social influences on the choice of a linguistic variant. In *Language and Culture in Society: A Reader in Linguistics and Anthropology*, D. Hymes & O. Werner (eds), 483–488. New York NY: Harper & Row.

Gauchat, L. 1905. L'unité phonetique dans le patois d'une commune. In *Aus Romanischen Sprachen and Literaturen: Festschrift Heinrich Mort*, L. P. Betz (ed.), 175–232. Halle: Max Niemeyer.

Goldberg, A. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: OUP.

Jakobson, R. 1971. *Studies on Child Language and Aphasia*. The Hague: Mouton. https://doi.org/10.1515/9783110889598

Johnstone, B. 1999. Uses of Southern speech by contemporary Texas women. *Journal of Sociolinguistics* 3: 505–522. https://doi.org/10.1111/1467-9481.00093

Johnstone, B. 2000. The individual voice in language. *Annual Review of Anthropology* 29: 405–424. https://doi.org/10.1146/annurev.anthro.29.1.405

Johnstone, B. & Bean, J. 1997. Self-expression and linguistic variation. *Language in Society* 26(2): 221–246. https://doi.org/10.1017/S0047404500020911

Labov, W. 1989. The exact description of the speech community: Short 'a' in Philadelphia. In *Language Change and Variation*, R. Fasold & D. Schiffrin (eds), 1–57. Washington DC: Georgetown University Press. https://doi.org/10.1075/cilt.52.02lab

Lamb, S. M. 1999. *Pathways of the Brain: The Neurocognitive Basis of Language*. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.170

Langacker, R. W. 1987. *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford CA: Stanford University Press.

Le Page, R. B. & Tabouret-Keller, A. 1985. *Acts of Identity: Creole-based Approaches to Language and Ethnicity*. Cambridge: CUP.

Milroy, L. 1987. *Language and Social Networks*. Oxford: Basil Blackwell.

Mollin, S. 2009. "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics* 14(3): 367–392. https://doi.org/10.1075/ijcl.14.3.04mol

Patrick, P. L. 2002. The speech community. In *Handbook of Language Variation and Change*, J. K. Chambers, P. Trudgill & N. Schilling-Estes (eds), 573–598. Oxford: Blackwell.

Pollard, C. & Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago IL: The University of Chicago Press.

Rayson, P. 2008. Wmatrix: a web-based corpus processing environment. Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/> (7 August 2017).

Sapir, E. 1973. *Selected Writings of Edward Sapir in Language, Culture, and Personality*. Berkeley CA: University of California Press

Saussure, F. 1985. The Linguistic Sign. In *Semiotics: An Introductory Anthology*, R. E. Innis (ed.), 24–46. Bloomington IN: Indiana University Press.

Talmy, L. 2000. *Toward a Cognitive Semantics*, Vol. 1. Cambridge MA: The MIT Press.

Teubert, W. 2010. *Meaning, Discourse and Society*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511770852

Tracking the third code

A cross-linguistic corpus-driven approach to metadiscursive markers

Sylviane Granger
Université catholique de Louvain

Corpus-driven methods are a powerful heuristic for identifying features of translated language, particularly phraseological features, which have been largely neglected in crosslinguistic studies. This chapter presents the results of an n-gram based study of metadiscursive markers in original and translated English. Automatic comparison of three- to four-word recurrent sequences in the two corpora reveals a series of markers that are significantly overused or underused in translated language. One category of overused bundles, i.e. markers of contrast, is analysed in the light of translation universals and source language effects. In a wider perspective, the study highlights the benefits of a multi-corpus empirical basis for corpus-based crosslinguistic studies and draws the preliminary contours of a methodological framework, 'Contrastive Translation Analysis', which integrates comparable, translation and learner corpus data.

1. Introduction

As demonstrated by a wide range of corpus-based cross-linguistic studies, translated language has its own distinctive features which set it apart from both source language (SL) and original target language (TL) texts. These features are often referred to as 'translationese', a term which has been given a range of definitions. It is often used in a clearly negative sense to refer to linguistic elements in the translated text that result from "the translator's inexperience or lack of competence" (Baker 1993: 249). For Rayson et al. (2008), translationese results in "mechanical and monotonous translated text at best, and inaccurate and erroneous translation in the worst cases". Gellerstam (1986: 88), on the other hand, uses the term in a more neutral meaning and explicitly excludes "anecdotal instances of bad translations". As the term 'translationese' tends to have negative undertones,

https://doi.org/10.1075/scl.87.08gra

I prefer to use Frawley's (2000 [1984]) non-evaluative 'third code'. For Frawley (2000: 257), translation "is essentially a third code which arises out of the bilateral consideration of the matrix and target codes"; it is "a code in its own right, setting its own standards and structural presuppositions and entailments, though they are necessarily derivative of the matrix information and target parameters". Although the terms 'translationese' and 'third code' are sometimes used interchangeably (see, for example, Balaskó 2008), they actually represent quite different perspectives. Shuttleworth and Cowie (1997: 173) insist that "while the two notions are clearly related, the term *third code* generally denotes more subtle deviations from TL linguistic norms, and its use implies on the part of the writer not only a lack of disapproval, but also the belief that such phenomena are worthy of systematic investigation for their own sake". The reality of translated language as a separate code is validated by the fact that it can be detected automatically. Baroni and Bernardini (2006) and Kurokawa, Goutte and Isabelle (2009), for example, report about 90% accuracy in detecting original vs. translated texts automatically on the basis of the frequency of words, lemmas, POS tags, etc. Koppel and Ordan (2011) go one step further: they report not only an accuracy rate of 96.7% in distinguishing original from translated texts, but also 92.7% accuracy in identifying the source language of a given translated text (see also Cartoni, Zufferey & Meyer 2013, Volansky, Ordan & Wintner 2013).

Studies of the third code in the related fields of contrastive linguistics (CL) and translation studies (TS) present both similarities and differences. The main similarity is that both fields rely on comparisons of corpora of original and translated texts in the same language to uncover manifestations of the third code. The differences reflect the respective objectives of each discipline. CL resorts to this type of comparison as one additional instrument to describe the systemic similarities and differences between languages: distinctive features of one language are identified through traces in their translation into one or more languages. In accordance with this objective, the corpus design of CL studies invariably includes the source texts of all translated texts. In his major opus on the use of corpora in contrastive studies, Johansson (2007: 33) makes it clear that general features of translated texts will not be discussed and that "[w]here translation effects are mentioned, the reference is usually to features induced by the source text".

In TS, comparisons of translated and non-translated texts contribute to one of the main objectives of the field, which is to gain a better understanding of the translation process by focusing on features of the translation product and, in particular, to uncover and test translation universals (Baker 1993, Mauranen & Kujamäki 2004), i.e. general laws and regularities of translation. As the focus is on "translation-based deviations from target language conventions" (Doherty 1998: 235) rather than

source-based deviations, the corpora used[1] typically only contain translated texts; the source texts are not included. The chapter devoted to features of translation in Olohan's 2004 book on corpora in translation studies purposely centres "not on source-text influence, but on other features of the translation process that may be reflected in the texts and language produced by translators" (Olohan 2004: 90–144).

This dual outlook on the same object of study can be of great mutual benefit to researchers in both fields: CL specialists can gain a heightened awareness of the possible impact of translation norms, while TS specialists are reminded of the possible impact of systemic differences between languages. As rightly observed by Ebeling and Oksefjell Ebeling (2013: 46), "contrastive linguistics and translation studies are two separate disciplines with their own theoretical foundations and research agendas. However, they are also closely related to each other, and performing one without the knowledge of theories and findings in the other may yield an incomplete or defective study."

The aim of this chapter is to shed fresh light on the notion of third code by drawing insights from both CL and TS with added insights from a related field, that of learner corpus research. Section 2 is devoted to the thorny issue of interpretation of third code effects. Section 3 makes the case for phraseology as a particularly promising locus of investigation into third code effects. Section 4 reports the preliminary results of a pilot study of an n-gram-based study of metadiscursive markers in original and translated English texts. The conclusion (Section 5) points to avenues for future research.

2. The issue of interpretation

It is one thing to uncover third code effects, quite another to interpret them. Teasing out universal and source effects has been a lingering problem for both contrastive linguists and translation specialists. For example, Espunya (2007) investigates V-*ing* adjuncts in translations from English to Catalan. She finds a sizeable proportion of explicitation by means of conjunctions and sentence connectives but is unable to attribute it to one of two possible causes: the translator's adjustment to Catalan, which has a preference for explicit linkage, or the explicitation universal.

Interesting insights on this issue can be gained from learner corpus research (LCR), a neighbouring field which centres around the analysis of foreign/second language learners on the basis of large corpora of written or spoken data (Granger

1. A good example is Baker's *Translational English Corpus* <https://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-english-corpus-tec/> (11 August 2018).

2012). There are striking similarities between TS and LCR. First, both involve interlingual mediation (Viaggio 2006), a status made clear by the term 'interlanguage' used in second language acquisition studies in general to refer to a language variety that has a structurally intermediate status between the native (L1) and target (L2) language. As pointed out by Chesterman (1998: 199), "Interlanguage studies, after all, are basically also contrastive, between the target language and the learner's interlanguage version of it". Second, both fields have undergone a similar evolution. The general outlook has moved from negative – with a heavy focus on errors and interference – to neutral. Third, studies of learner language and translated language have adopted an increasingly corpus-driven methodology. Finally, both fields are confronted with the same difficulty, i.e. teasing out L1 effects from general features of acquisition for LCR, and SL effects from translation universals for TS. Interestingly, this issue is characterised by the same pendulum swing: the dominant focus on universal explanations is progressively being replaced by a more balanced view which also leaves room for specific (i.e. L1-/SL-related) explanations.

In view of these similarities several researchers have called for a mutually beneficial rapprochement. Johansson (2007: 313) observes that "[n]ew possibilities are afforded by the combined use of learner corpora and multilingual corpora". Chesterman (2007: 62–63) echoes this point from a TS perspective: "We could add to our reference texts corpora representing target-language production in a different set of special conditions: texts produced by non-native writers. I will call these LT, for learners' texts. After all, both translations and learners' texts are produced under particular constraints, and it may be that these constraints have similar effects". A good illustration of this approach is provided by Gaspari and Bernardini (2010), who have compiled a monolingual comparable corpus of non-native and translated language and analyse it to explore the hypothesis that the two types of language "share similar features, and that these can be accounted for by the notion 'mediation corpora' in search of mediation rather than translation universals" (Gaspari & Bernardini 2010: 228). One major difference between the two types of data, however, is that translators usually translate into their mother tongue, while learners write in a foreign or second language.

In the early days of LCR I also argued in favour of greater synergy between corpus-based cross-linguistic and acquisition studies: "Far from seeing computerized bilingual corpora as the private ground of translation specialists and typologists, and computerized learner corpora as the sole concern of applied linguists, we see the two types of corpora as closely interrelated" (Granger 1996: 46). The methodological framework I put forward, the 'Integrated Contrastive Model', integrates Contrastive Analysis (CA), which compares different languages, and a new type of contrastive analysis, 'Contrastive Interlanguage Analysis' (CIA)

(Granger 1996, 2015), which compares learner and native (or expert) varieties of the same language. The model relies on a wide range of corpora: for CA, comparable corpora of original texts in different languages and translation corpora made up of source texts in one language and translations in another; for CIA, corpora representing different interlanguage varieties, for example texts written by French and Chinese learners, compared with each other and/or with one or more native (or expert) reference language varieties. The aim of the model is to contribute to the identification of transfer effects. Bilingual corpora representing the learner's mother tongue and target language can be used to predict the occurrence of some linguistic features in learner data. Alternatively, distinctive features of learner language identified by CIA can be diagnosed as potentially transfer-related by a cross-linguistic study of these features in bilingual corpora of the two languages involved. One element that greatly helps the interpretation is the fact that more than one learner population is investigated. Comparing the same feature in different learner populations provides valuable information on the degree of generalisability of the features. A corpus like the *International Corpus of Learner English* (Granger et al. 2009), which contains data from sixteen different mother tongue backgrounds, is a particularly useful resource to tease out developmental vs. L1-specific features of interlanguage. Linguistic features are regularly presented as transfer-related in studies based on learner corpus data from one particular learner population when, in fact, they are also found in several other learner populations and are therefore at least partly developmental.

A similar model adapted to a TS perspective could contribute to strengthening the empirical basis of the field and offer a partial answer to De Sutter et al.'s (2012) general call for greater methodological rigour in corpus-based translation studies. Figure 1 is a first attempt at representing the corpus design with English as the focal language.[2] In parallel with the CIA term, I propose to refer to the approach as 'Contrastive Translation Analysis' (CTA). The main rationale behind the corpus design is that "the study of translation cannot be performed successfully without acknowledging source structures, texts or languages" (De Sutter et al. 2012: 142). In keeping with this basic tenet, the corpus base includes large corpora of translated language, which can be used as wholes, i.e. with no reference to the source texts, but can also be broken down into SL-specific subcorpora, thereby allowing pairwise comparisons between source and translated texts. It also contains large comparable corpora of original, i.e. non-translated texts, in all the languages involved in the analysis. The model also includes learner corpora, broken down according to the learners' mother tongue background, thereby answering the call

2. The full CTA model is bidirectional. The other translation direction (from English to French/Dutch/Chinese) is not shown to make the figure easier to read.

for a rapprochement between cross-linguistic and acquisition studies.[3] As regards the analysis of third code features, the rationale for including learner corpus data is that the presence of the same feature in, say, English translations of French texts and English texts written by French learners, constitutes additional confirmatory evidence for the presence of source language effects.

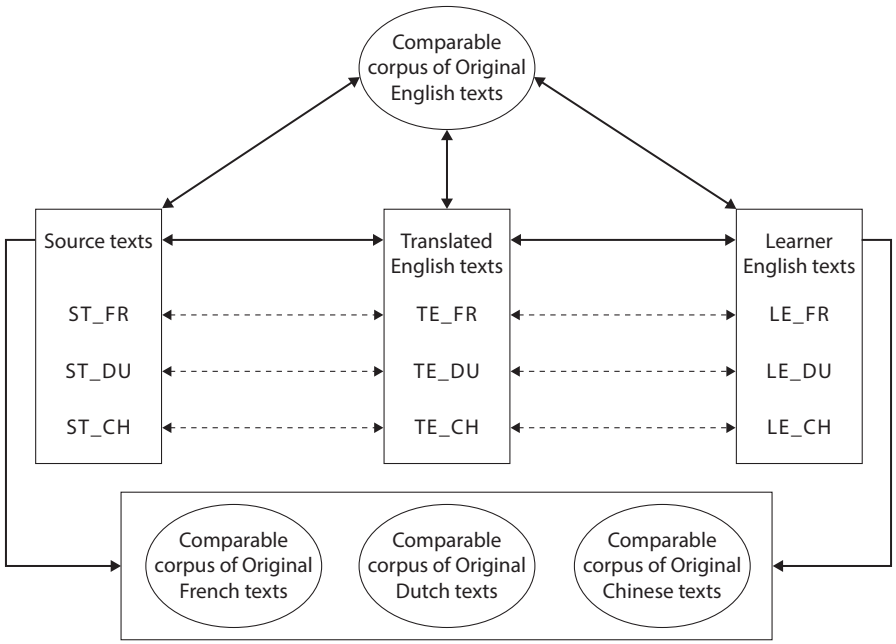

Figure 1. Contrastive Translation Analysis (CTA)

3. Phraseology: A prime locus for the third code

The third code manifests itself in a highly diversified set of linguistic features: morphological (e.g. Lefer & Cartoni 2013), grammatical (e.g. Dai & Xiao 2011), lexical (e.g. Frankenberg-Garcia 2008) and discursive (e.g. Pisanski Peterlin 2010). One aspect that has been comparatively neglected is phraseology, i.e. the study of word combinations. Cross-linguistic studies that have investigated phraseological aspects of language have tended to focus on semantically non-compositional, mainly figurative, units. In CL, Piirainen (2008) and Ishida (2008) are examples of studies that focus on idioms in the strict sense, i.e. word combinations characterised by a

3. Note that the ideal situation represented in Figure 1 will be very difficult to achieve for smaller languages that have a limited number of corpus resources.

high degree of formal, syntactic and semantic fixedness, such as *to be in the red* or *to fly off the handle*. In TS, Baker (2007) focuses on idioms that are characterised by a high degree of opacity such as *off the hook* in translated vs. non-translated English. The study suggests that such idioms tend to be underused in translated English and, when present, are more likely to be used in their literal rather than idiomatic meaning.

With the advent of corpus linguistics, however, the scope of phraseology has considerably widened. Inspired by John Sinclair's ground-breaking work, the corpus-based approach to phraseology has opened up a "huge area of syntagmatic prospection" (Sinclair 2004: 19). Using automatic extraction methods, researchers have identified a variety of lexical co-occurrences, which, unlike idioms, are highly frequent and tend to be situated towards the semantically compositional side of the transparency-opacity continuum. One method that has become highly popular, especially in variationist (e.g. Conrad & Biber 2004) and second language acquisition studies (e.g. Chen & Baker 2010) is n-gram extraction. This method consists in extracting recurrent sequences of contiguous words (two words, three words, etc.) from corpora. Admittedly very crude linguistically, it has proved to be a very powerful discovery tool, capable of uncovering a wide range of pre-packaged expressions that had hitherto been neglected. It is this method that underlies Biber et al.'s 'lexical bundles', i.e. "sequences of words that most commonly co-occur in a register" (1999: 989). Biber, Conrad and Cortes (2004) divide these sequences into three main categories according to their function in discourse: (1) referential bundles, which make direct reference to physical or abstract entities, or to the textual context itself (*a lot of people, in the United States*); (2) discourse organizers, which reflect relationships between prior and coming discourse (*with this in mind, this is why*); and (3) stance bundles, which express attitude or assessment of certainty (*I don't want to, it is possible to*). These sequences are very important for foreign language learners (and, one might add, for translators), as "producing natural, idiomatic English is not just a matter of constructing well-formed sentences, but of using well-tried lexical expressions in appropriate places" (Biber et al. 1999: 990).

In cross-linguistic studies the n-gram method has only recently started to be used. Some researchers have extracted and compared lexical bundles in different languages: English and Spanish (Cortes 2008), English and Italian (Forchini & Murphy 2008), English, Spanish and Korean (Biber, Kim & Tracy-Ventura 2010), English and Norwegian (Ebeling & Oksefjell Ebeling 2013) and English and French (Granger 2014). These studies brought to light some interesting cross-linguistic differences: for example, Cortes (2008) observed that Spanish had a greater tendency than English to use organisational markers to introduce topics or mark focus, while English used them more than Spanish to elaborate the message.

Other researchers have compared the frequency and use of lexical bundles in translated and non-translated texts in the same language. In an article presented as "an exercise in methodology", Baker (2004) used the n-gram method to pull out lists of phrases of various lengths from corpora of translated and non-translated narratives. She focused on the phrases that displayed different frequencies in the two corpora (such as *that is to say* or *in other words*), excluding those that were clearly tied to the theme of a single text (i.e. referential bundles in Biber et al.'s classification). In her analysis, Baker purposely disregards the potential influence of source texts and focuses on the impact of individual translators' preferences. Similar work by Xiao (2011) based on corpora of original and translated Chinese showed that word clusters of all types were much more frequent in translated Chinese, a tendency which he attributes in part to the influence of the English source texts. Lee (2012), on the other hand, compares Korean texts translated from English and non-translated Korean journalistic texts. Like Baker (2004), Lee restricts his analysis to stance and organisational bundles. One particularly interesting result of the study is the fact that hedging expressions are under-represented in Korean translations, thereby giving translated Korean an assertive tone that is atypical of original Korean texts.

4. Metadiscursive markers in original and translated English

The few studies reported above suggest that the n-gram method can play a key role in the extraction and analysis of a wide range of metadiscursive markers, i.e. linguistic items that do not add propositional material but help readers organise, classify, interpret, evaluate and react to such material (Vande Kopple 1985). This is particularly welcome news for translation studies, as metadiscourse has been relatively neglected. Nord (2007) observes that phatic aspects of language in general have not been given much attention in translation studies and that "there is still plenty of gold in the mine" (Nord 2007: 183). N-gram extraction is a promising method to mine that gold. For one thing, metadiscourse is often expressed by longer sequences; these have been largely disregarded in cross-linguistic studies and can easily be extracted using the n-gram technique. More importantly, the method is completely corpus-driven and therefore capable of detecting a much wider range of markers than researchers could ever hope to identify on the sole basis of their intuition or small data samples.

In this section I will present some preliminary results of an n-gram-based study of metadiscursive markers in original and translated English. In keeping with the CTA framework, sources are identified and my main focus will be on English translated from French. This language pair is fertile ground for the study

of third code effects, as English and French have traditionally been presented as displaying marked metadiscursive differences. French is described as more explicitly conjunctive and emphatic than English (Vinay & Darbelnet 1995 [1958]: 234ff and 220ff, Delisle 1993: 432), and generally more verbose. It can therefore be assumed that metadiscursive bundles will tend to be more frequent in English translated from French than in original English texts.

4.1 Data and methodology

The corpus used is a version of the *Europarl* corpus with clearly identified original vs. translated texts (Cartoni, Zufferey & Meyer 2013).[4] The subcorpora used contain proceedings of the European parliament debates dated 1996–1999. Table 1 gives the breakdown of the different subcorpora. The main focus of the study will be on English original texts and English texts translated from French. To fine-tune the interpretation of the results, three other corpora will be used: English texts translated from Dutch, as well as French and Dutch originals.

Table 1. Breakdown of *Europarl* corpora

***Europarl* corpora**	**Total number of words**
English original texts (OE)	1,381,218
English texts translated from French (TE_FR)	1,169,081
English texts translated from Dutch (TE_DU)	831,081
French source texts (FR)	1,317,774
Dutch source texts (DU)	830,051

The methodology comprises three main stages, the first two being fully automatic. First, all recurrent sequences of three to four words are extracted from the original English (OE) texts and the English texts translated from French (TE_FR). This is done via the Wordlist function of *WordSmith Tools* (Scott 2008). Second, the generated lists of bundles in OE and TE_FR are automatically compared to extract the key bundles, i.e. the bundles that are significantly more or less frequent in TE_FR than in OE. This is done using *WordSmith Tools*' keyword function which allows comparison of lists of both single words and clusters. The statistical test used is log-likelihood (LL) and the p value was set at 0.001. The third stage is manual: all of the referential bundles are discarded so as to keep only the two metadiscursive categories, i.e. Biber, Conrad and Cortes's (2004) discourse organizers and stance bundles.

4. I would like to express my gratitude to Bruno Cartoni for giving me access to this version of the corpus.

4.2 Over- and underused metadiscursive markers

The keyword technique is a very powerful heuristic: it brings out a wealth of potentially interesting strings to explore. A sample of significantly over- and underused bundles is included in Table 2. Underused bundles contain a very large proportion of bundles with the personal pronoun *I*, a feature which is most probably prompted by the French source texts. In a previous study, likewise based on *Europarl* (Granger 2014), I established that this type of bundle was much more frequent in English than in French and generally related to a greater preference for the use of first-person pronouns in English (cf. Vassileva 1998, Breivega, Dahl & Fløttum 2002).

Table 2. Sample of over- and underused lexical bundles in TE_FR (raw frequencies)

Underused lexical bundles	OE	TE_FR	LL	Overused lexical bundles	OE	TE_FR	LL
I WANT TO	567	81	334.2	IN ORDER TO	310	926	431.7
I WISH TO	386	84	163.6	THIS IS WHY	24	212	204.7
I AM SURE	321	62	151.6	IN THIS RESPECT	56	266	185.8
IT IS IMPORTANT THAT	159	20	101.1	I THINK THAT	117	322	136.5
I HOPE THAT	414	157	81.4	WITH REGARD TO	289	551	132.2
I HAVE TO SAY	168	36	72.3	WITHIN THE FRAMEWORK OF	16	136	129.3
IN RELATION TO	345	132	66.6	IN OTHER WORDS	50	197	119.6
I UNDERSTAND THAT	82	6	66.2	ON THE OTHER HAND	59	198	104.1
WE NEED TO	473	212	63.5	WHAT IS MORE	4	74	88.7
I WOULD ASK	157	36	63.1	AND ABOVE ALL	15	98	82.7
I WONDER WHETHER	58	4	47.8	AS REGARDS THE	60	178	82.3
I WISH TO BEGIN	50	3	43.0	ON THE CONTRARY	17	99	78.5
IT IS QUITE CLEAR	49	3	41.9	IN ANY CASE	28	118	75.6
WE WANT TO SEE	75	12	41.0	IT IS TRUE	39	131	68.9
I JUST WISH	36	1	36.6	IN MY OPINION	42	132	65.0
WE WANT TO	237	101	36.1	IT IS TRUE THAT	28	106	62.2
IT IS VERY IMPORTANT	74	14	35.6	AT A TIME WHEN	39	119	56.8
IN THAT REGARD	28	0	34.4	ALL THE MORE	21	87	55.0
I JUST WISH TO	33	1	33.1	AM THINKING IN PARTICULAR	1	39	52.7

Table 2. (*continued*)

Underused lexical bundles	OE	TE_FR	LL	Overused lexical bundles	OE	TE_FR	LL
I WOULD HOPE	52	7	31.8	FOR MY PART	5	50	50.6
I AM GLAD	77	18	30.4	WE MUST THEREFORE	6	51	48.5
I VERY MUCH WELCOME	34	2	29.4	HOW CAN WE	24	85	47.0
I APPRECIATE THAT	23	0	28.2	I BELIEVE THAT	246	361	45.1
IN THIS REGARD	97	30	27.0	IN TERMS OF	335	457	44.5
CAN I SAY	31	2	26.1	ON THE ONE HAND	58	131	42.3
I WOULD HOPE THAT	40	5	25.5	AT ALL EVENTS	0	27	42.1
I VERY MUCH HOPE	33	3	24.5	AS FOR THE	20	72	40.4
WE MUST ENSURE THAT	62	15	23.6	ON THIS SUBJECT	32	90	39.1
I AM CONFIDENT THAT	32	3	23.5	CONTRARY TO THE	18	67	38.7
I SUSPECT THAT	19	0	23.3	IN OUR OPINION	5	39	35.8
I AM CONFIDENT	40	6	22.8	OF COURSE BUT	1	27	34.7
I WONDER IF	47	9	22.4	I WOULD REMIND YOU	6	38	31.5
FOR THAT REASON	36	5	21.6	I THEREFORE BELIEVE	0	20	31.2
WE NEED TO ENSURE	36	5	21.6	DUE TO THE	37	88	30.7
THERE SEEMS TO BE	22	1	20.3	I AM THINKING OF	2	27	30.0
WE ARE TALKING ABOUT	73	23	19.8	WHAT IS AT STAKE	2	27	30.0
I WANT TO SAY	46	10	19.5	LET US NOT	52	106	28.8
AS I UNDERSTAND IT	21	1	19.2	THE MORE SO	2	26	28.6
IT IS ABOUT TIME	21	1	19.2	FOR OUR PART	3	28	27.6
COULD I ASK	20	1	18.1	FOR MY PART I	5	32	26.7

Although the list of overused bundles contains a few with first-person pronouns, it is mainly characterised by a large number of (mainly) adverbial phrases or clauses that have an organisational or stance function. The following four categories stand out:

- Markers of contrast (*on the contrary, on the one hand, on the other hand*)
- Topicalizers (*as regards, with regard to, as far as x is concerned, as for*)
- Stance markers (*in fact, in actual fact, in point of fact, in reality, in truth*)
- *Let us* imperatives (*let us hope, let us not forget*)

In the next two sections I will focus on one category of overused bundles: markers of contrast. Logical connectors are particularly fertile ground for studying the third code, as they have been shown to be both susceptible to explicitation and prone to SL influence.

4.3 Translation universal or systemic cross-linguistic difference?

Table 3 gives the relative frequency per million words (pmw) of three significantly overused markers of contrast in English texts translated from French as compared to English original texts. This overuse could be due to the translation universal of explicitation. According to Blum-Kulka (1986[2000]: 300), the process of interpretation performed by the translator often leads to a higher degree of redundancy in the target text and this redundancy "can be expressed by a rise in the level of cohesive explicitness in the TL text". Several studies have supported the explicitation hypothesis. Xiao (2011), for example, reports much more frequent use of reformulation markers in translated Chinese than in original Chinese texts. A comparison with the source texts showed that 10–30% of the reformulation markers were absent from the source texts and added by the translator.

However, the frequency differences between OE and TE_FR could also be due to SL effects. Becher's 2011 study of connectives in a bidirectional corpus of English and German texts highlights a much higher degree of cohesive explicitness in German translations from English than in English translations from German. For Becher (2011: 29–30), "we do not need the assumption of a mysterious translation-inherent type of explicitation". In his view, most instances of explicitation are language pair-specific: they result from cross-linguistic differences between the languages involved.

We have good reason to expect the overuse evidenced by our data to result from a systemic difference between English and French. It is generally accepted that French uses more rhetorical devices than English and that "[t]heir literal translation into English can cause a feeling of incongruity because English would use them more sparingly". Similarly, Armstrong (2005: 196) observes that "the general tendency seems to be to mark cohesion in French in a more explicit way than in English, using more linguistic material". It is important to note, however,

Table 3. Frequency of markers of contrast in OE and TE_FR (rel.freq./pmw)

Markers of contrast	OE	TE_FR
on the contrary	12.3	84.6
on the one hand	41.9	112
on the other hand	42.7	169.3

that this systemic difference does not rest on a solid empirical foundation and therefore remains largely hypothetical.

The first step that can be taken with a view to teasing out the two possible interpretations is to compare TE_FR with texts translated from other languages. In this preliminary study only one other SL will be investigated, namely Dutch (TE_DU). As Table 4 shows, TE_DU occupies an intermediate position between OE and TE_FR. There is also a tendency to overuse the three markers of contrast but it is much less marked than in TE_FR. At this stage, the explicitation hypothesis cannot be discarded but the different results for TE_DU and TE_FR suggest that SL effects also play a role.

Table 4. Markers of contrast in OE, TE_DU and TE_FR (rel.freq./pmw)

Markers of contrast	OE	TE_DU	TE_FR
on the contrary	12.3	43.3	84.6
on the one hand	41.9	80.6	112
on the other hand	42.7	117.9	169.3

The impact of the explicitation hypothesis can be further investigated by means of a comparison of the three markers and their sources in the French and Dutch texts. As shown in Tables 5 and 6, the proportion of explicitation, i.e. the absence of a marker of contrast in the original text and the addition of one in the translated text, is minimal: 1.4% for French and 4.4% for Dutch.

Table 5. TE_FR vs. French source texts

TE_FR	FR_source texts
on the contrary (99)	*au contraire* (78), *bien au contraire* (9), *tout au contraire* (4), *à l'inverse* (3), *par contre* (2), other (2), **zero (1)**
on the one hand (131)	*d'une part* (84), *d'un côté* (37), *à la fois* (2), *pour une part* (2), other (3), **zero (3)**
on the other hand (198)	*en revanche* (69), *d'autre part* (58), *par contre* (21), *d'un autre côté* (13), *de l'autre* (10), *au contraire* (6), PRONOUN (*lui, elle(s), eux*) (4), *quant à* + PRONOUN (3), *de l'autre côté* (2), other (10), **zero (2)**
	Total explicitation: 6/428 = 1.4%

These figures are much lower than those found in other studies of logical connectives. As mentioned above, Xiao (2011) reports a much higher rate (10–30%), with peaks as high as 85% for some markers. Similarly, Cartoni et al. (2011) report high explicitation rates for causal connectives in English and French. Like our study, theirs is based on the *Europarl* corpus, but it has the added interest of investigating

Table 6. TE_DU vs. Dutch source texts

TE_DU	DU_source texts
on the contrary (36)	*integendeel* (31), other (2), **zero (3)**
on the one hand (67)	*enerzijds* (38), *aan de ene kant* (23), *langs de ene kant* (2), other (1), **zero (3)**
on the other hand (98)	*aan de andere kant* (36), *anderzijds* (36), *daarentegen* (11), *(maar) van de andere kant* (3), *langs de andere kant* (3), *tegenover* (2), other (4), **zero (3)**
	Total explicitation: 9/201 = 4.4%

both translation directions. This allows them to establish that explicitation rates are much higher when French is the target language than in the other direction.

The difference between these results and ours may partly be due to the type of cohesive linking. Xiao (2011) and Cartoni et al. (2011) have investigated reformulation and cause respectively; these relations have a weaker degree of informativeness (Kortmann 1991) and are therefore potentially more susceptible to explicitation than a strong relation like contrast. Another factor that may account for the minimal explicitation displayed by our data is the respective degree of cohesiveness of the two languages involved. It is reasonable to expect less explicitation when passing from a more explicitly cohesive language like French to a less explicitly cohesive language like English. As shown by Becher (2011), the opposite situation holds for the German-English pair.

At this stage, the balance seems to tip more in the direction of SL effects than explicitation. Additional insights can be gained from an analysis of learner corpus data. Using the CIA framework described in Section 1, I compared the frequency of the three markers in two corpora of argumentative writing: the *Louvain Corpus of Native English Essays* (LOCNESS),[5] which contains essays written by native English students, and the *International Corpus of Learner English* (ICLE) (Granger et al. 2009), which contains comparable texts written by higher intermediate to advanced learners of English. Table 7 presents the results of this comparison for two learner populations: Dutch-speaking learners (ICLE_DU) and French-speaking learners (ICLE_FR).[6]

5. <http://www.uclouvain.be/en/research-institutes/ilc/cecl/locness.html (11 August 2018)

6. The respective sizes of the corpora are: 149,627 words (LOCNESS), 227,613 words (ICLE_DU) and 191,077 words (ICLE_FR).

Table 7. Markers of contrast in native and learner English (rel.freq./pmw)

Markers of contrast	LOCNESS	ICLE_DU	ICLE_FR
on the contrary	13.3	70.3	303.5
on the one hand	0	96.6	151.7
on the other hand	147	294.3	376.8

A comparison between translated English (Table 4) and learner English (Table 7) highlights a much higher frequency of the markers in learner English texts. This difference most probably results from the difference in text type between the two corpora and is therefore of no particular interest for our study. Apart from this difference, the two sets of results are remarkably similar: in both cases Dutch occupies an intermediate position between the native English texts and the 'mediated' texts. This finding reinforces the conclusion that SL effects are at play in the two languages, albeit more strongly for French than for Dutch.

5. Conclusion

While third code effects are now much easier to identify thanks to powerful corpus methods, they remain very difficult to interpret. The view advanced here is that significant benefits can be derived from a multi-disciplinary approach, combining the strengths of contrastive linguistics, translation studies and learner corpus research. To achieve its full effect, this approach needs to rest on a solid, multi-corpus empirical basis. The CTA design is one preliminary step towards a more comprehensive model for corpus-based cross-linguistic studies which integrates all the relevant types of data, shows how they interact and specifies which types of corpus data are needed for which type of investigation.

An approach along these lines can greatly contribute to a more rigorous evaluation of the respective impact of translation universals and source language effects. For the three connectors of contrast analysed in this study, the main effect appears to be the frequency of this type of cohesive marker in the source language; there is very little evidence of explicitation. This does not disprove the explicitation hypothesis but suggests that the explicitation rate is a function of the respective degree of cohesiveness of the two languages concerned and the type of linking involved, some types of links being more prone to explicitation than others. Though admittedly limited, our results are in line with Volansky, Ordan and Wintner's (2013: 6) conclusion that universal claims "vary greatly across different pairs of languages" and lend support to their call for a reconsideration of the notion of translation universal.

Like all corpus-driven methods, the *n*-gram method is a powerful heuristic with great potential for cross-linguistic studies. The corpus-based/corpus-driven distinction (Tognini-Bonelli 2000) highlights the difference between studies that exemplify or test existing theories and studies that use data as a starting point to formulate new theories. Olohan (2004: 190) observes that translation studies have tended to be predominantly corpus-based and would benefit from more corpus-driven methods. The extraction of *n*-grams is one such method. Our study shows that it provides a unique window onto pragmatics and rhetoric and can help confirm or disconfirm sweeping generalisations about languages which, though credible and probably well founded (or at least partly), are still awaiting empirical validation.

It needs to be pointed out, however, that the current study suffers from one important methodological limitation. The statistical measures underlying the keyword analysis have been computed on the basis of aggregate data, i.e. corpora taken as wholes without taking any account of variability. This is problematic, as both learners and translators can have their own preferences for particular cohesive markers or types of cohesive markers and this can influence and potentially skew the results. The only way of avoiding this pitfall is to use statistical tests that take variability into account. Paquot (2014), for example, uses Wilcoxon rank sum tests to identify and analyse transfer effects in French learners' use of lexical bundles in the ICLE_FR compared to nine other ICLE subcorpora. Although one cannot expect complete overlap between Paquot's results and those reported here, in view of differences in statistical techniques and data selection criteria, it is interesting to note that many of the bundles she identifies as L1-related are the same as those we have identified as being distinctive of English texts translated from French, such as *on the contrary, as far as * is concerned* or *let us* imperatives.

One important issue that has been left out of the discussion so far is that of the practical use made of third code effects once they have been detected. The answer to this question clearly depends on how close to source language one wishes translated language to be. As pointed out by Frawley (2000 [1984]: 260), this is a matter of preference: "Evaluative discussions on recodification are matters of preference solely. Consider, in this regard, the fact that the fidelity of a new linguistic text to its 'original' is often viewed as the criterion of goodness for interlingual translation. But is the 'original' text the matrix or the target code? Each contributes to the genesis of the translation." Taking the viewpoint of the lexicographer, Teubert (1996: 247) is extremely critical of translations which "however good and near-perfect they may be (but rarely are), cannot but give a distorted picture of the language they represent". From this perspective, third code effects are clearly negative and indeed, Teubert argues against the use of translation corpora for lexicographic purposes. A much more positive stand is taken by Wakabayashi

(2009), who explains that in Japan there is a generally positive attitude towards a "foreign-tinged style in translations into Japanese", which "contrasts with the inward-looking expectation in Anglophone circles that translations should sound smooth and natural in the target language". But this mainly concerns literary translation, a field where target culture orientation is a key strategy. In many other contexts, in particular that of translator training, the situation is quite different. It is useful to draw students' attention to third code features, especially those that give a text an unwanted translated feel. Phraseologisms rank highly among these features and, as pointed out by Colson (2008: 201–202), they could play a key role in automatic translation assessment. In both theoretical and practical terms, there is no doubt that phraseology in the wide sense, and metadiscursive phraseology in particular, holds great promise for cross-linguistic studies and should therefore be the focus of greater attention in future research.

Acknowledgement

The research reported here was first presented as a plenary talk given at the ICLC 7 – UCCTS 3 Conference held at Ghent University, Belgium, 11–13 July 2013.

References

Armstrong, N. 2005. *Translation, Linguistics, Culture: A French-English Handbook*. Clevedon: Multilingual Matters.

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 233–250. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.15bak

Baker, M. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2): 167–193. https://doi.org/10.1075/ijcl.9.2.02bak

Baker, M. 2007. Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics* 21: 11–21. https://doi.org/10.1075/bjl.21.02bak

Balaskó, M. 2008. What does the *figure* show? Patterns of translationese in a Hungarian comparable corpus. *Trans-kom* 1(1): 58–73.

Baroni, M. & Bernardini, S. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3): 259–274. https://doi.org/10.1093/llc/fqi039

Becher, V. 2011. When and why do translators add connectives? A corpus-based study. *Target* 23(1), 26–47. https://doi.org/10.1075/target.23.1.02bec

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

Biber, D., Conrad, S. & Cortes, V. 2004. If you look at … Lexical bundles in university lectures and textbooks. *Applied Linguistics* 25: 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Kim, Y-J. & Tracy-Ventura, N. 2010. A corpus-driven approach to comparative phraseology: Lexical bundles in English, Spanish, and Korean. In *Japanese and Korean Linguistics*, S. Iwasaki, H. Hoji, P. Clancy & S-O. Sohn (eds), 75–94. Stanford CA: CSLI.

Blum-Kulka, S. 1986[2000]. Shifts of cohesion and coherence in translation. In *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, J. House & S. Blum-Kulka (eds), 17–35. Tübingen: Narr. (Reprinted in Venuti, L. (ed.). 2000. *The Translation Studies Reader*, 298–313. London: Routledge).

Breivega K., Dahl, D. & Fløttum, K. 2002. Traces of self and others in research articles. A comparative pilot study of English, French and Norwegian research articles in medicine, economics and linguistics. *International Journal of Applied Linguistics* 12(2): 218–239. https://doi.org/10.1111/1473-4192.00032

Cartoni, B., Zufferey, S., Meyer, T. & Popescu-Belis, A. 2011. How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, Portland, Oregon, 24 June, 78–86. Stroudsburg PA: ACL.

Cartoni, B., Zufferey, S. & Meyer, T. 2013. Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics* 27: 23–42. https://doi.org/10.1075/bjl.27.02car

Chen, Y-H. & Baker, P. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14 (2): 30–49.

Chesterman, A. 1998. *Contrastive Functional Analysis* [Pragmatics & Beyond New Series 47]. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.47

Chesterman, A. 2007. Similarity analysis and the translation profile. *Belgian Journal of Linguistics* 21: 53–66. https://doi.org/10.1075/bjl.21.05che

Colson, J. 2008. Cross-linguistic phraseological studies. An overview. In *Phraseology: An Interdisciplinary Perspective*, S. Granger & F. Meunier (eds), 191–206. Amsterdam: John Benjamins. https://doi.org/10.1075/z.139.19col

Conrad, S. & Biber, D. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* 20: 56–71.

Cortes, V. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3(1): 43–57. https://doi.org/10.3366/E1749503208000063

Dai, G. & Xiao, R. 2011. 'SL shining through' in translational language: A corpus-based study of Chinese translation of English passives. *Foreign Language Teaching Theory and Practice* 1: 8–15.

Delisle, J. 1993. *La traduction raisonnée. Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*. Ottawa: Presses de l'Université d'Ottawa.

De Sutter, G., Goethals, P., Leuschner, T. & Vandepitte, S. 2012. Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures* 13(2): 137–143. https://doi.org/10.1556/Acr.13.2012.2.1

Doherty, M. 1998. Clauses or phrases – A principled account of *when*-clauses in translations between English and German. In *Corpora and Cross-Linguistic Research*, S. Johansson & S. Oksefjell Ebeling (eds), 235–254. Amsterdam: Rodopi.

Ebeling, J. & Oksefjell Ebeling, S. 2013. *Patterns in Contrast* [Studies in Corpus Linguistics 57]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.58

Espunya, A. 2007. Is explicitation in translation cognitively related to linguistic explicitness? *Belgian Journal of Linguistics* 21: 67–86. https://doi.org/10.1075/bjl.21.06esp

Forchini, P. & Murphy, A. 2008. N-grams in comparable specialized corpora. Perspectives on phraseology, translation, and pedagogy. *International Journal of Corpus Linguistics* 13(3): 351–367. https://doi.org/10.1075/ijcl.13.3.06for

Frankenberg-Garcia, A. 2008. 'Suggesting rather special facts': A corpus-based study of distinctive lexical distributions in translated texts. *Corpora* 3(2): 195–211. https://doi.org/10.3366/E1749503208000154

Frawley, W. 2000[1984]. Reprint of 'Prolegomenon to a theory of translation'. In *The Translation Studies Reader*, L. Venuti (ed.), 250–263. London: Routledge.

Gaspari, F. & Bernardini, S. 2010. Comparing non-native and translated language: Monolingual comparable corpora with a twist. In *Using Corpora in Contrastive and Translation Studies*, R. Xiao (ed.), 215–234. Newcastle upon Tyne: Cambridge Scholars.

Gellerstam, M. 1986. Translationese in Swedish novels translated from English. In *Translation Studies in Scandinavia*, L. Wollin & H. Lindquist (eds), 88–95. Lund: CWK Gleerup.

Granger, S. 1996. From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In *Languages in Contrast. Text-based Cross-linguistic Studies*, K. Aijmer, B. Altenberg & M. Johansson (eds), 37–51. Lund: Lund University Press.

Granger, S. 2012. Learner corpora. In *The Encyclopedia of Applied Linguistics*, C. A. Chapelle (ed.), 3235–3242. Oxford: Wiley-Blackwell.

Granger, S. 2014. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14(1): 58–72. https://doi.org/10.1075/lic.14.1.04gra

Granger, S. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1): 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *The International Corpus of Learner English*. Handbook and CD-ROM, Version 2. Louvain-la-Neuve: Presses universitaires de Louvain.

Ishida, P. 2008. Contrastive idiom analysis: The case of Japanese and English idioms of anger. In *Phraseology: An Interdisciplinary Perspective*, S. Granger & F. Meunier (eds), 275–291. Amsterdam: John Benjamins. https://doi.org/10.1075/z.139.24ish

Johansson, S. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies* [Studies in Corpus Linguistics 26]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.26

Koppel, M. & Ordan, N. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, 1318–1326. Portland OR: ACL.

Kortmann, B. 1991. *Free Adjuncts and Absolutes in English. Problems of Control and Interpretation*. London: Routledge.

Kurokawa, D., Goutte, C. & Isabelle, P. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of the Twelfth Machine Translation Summit* hosted by the Association for Machine Translation in the Americas. <http://www.mt-archive.info/MTS-2009-Kurokawa.pdf> (11 August 2018).

Lee, C. 2012. Using lexical bundle analysis as discovery tool for corpus-based translation research. *Studies in Translation Theory and Practice* 21(3): 1–18.

Lefer, M-A. & Cartoni, B. 2013. Word-formation in original and translated English: Source language influence on the use of *un-* and *-less*. *Quaderns de Filologia, EstudisLinguistics* XVIII: 49–59.

Mauranen, A. & Kujamäki, P. (eds). 2004. *Translation Universals. Do they exist?* [Benjamins Translation Library 48]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.48

Nord, C. 2007. The phatic function in translation: Metacommunication as a case in point. *Belgian Journal of Linguistics* 21: 171–184. https://doi.org/10.1075/bjl.21.12nor

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.

Paquot, M. 2014. Cross-linguistic influence and formulaic language. Recurrent word sequences in French learner writing. *EUROSLA Yearbook* 14: 240–261. https://doi.org/10.1075/eurosla.14.10paq

Piirainen, E. 2008. Phraseology in a European framework: A cross-linguistic and cross-cultural research project on widespread idioms. In *Phraseology: An Interdisciplinary Perspective*, S. Granger & F. Meunier (eds), 243–258. Amsterdam: John Benjamins. https://doi.org/10.1075/z.139.22pii

Pisanski Peterlin, A. 2010. Hedging devices in Slovene-English translation: A corpus-based study. *Nordic Journal of English Studies* 9(2): 171–193.

Rayson, P., Xu, X., Xiao, J., Wong, A. & Yuan, Q. 2008. Quantitative analysis of translation revision: Contrastive corpus research on native English and Chinese translationese. In *Proceedings of XVIII FIT World Congress*, Shanghai, China, 4–7 August, 2008. <http://eprints.lancs.ac.uk/41883/1/Rayson_P_Et_Al_fit2008.pdf (11 August 2018).

Scott, M. 2008. *WordSmith Tools*, Version 5. Liverpool: Lexical Analysis Software.

Shuttleworth, M. & Cowie, M. 1997. *Dictionary of Translation Studies*. London: Routledge.

Sinclair J. 2004. *Trust the Text – Language, Corpus and Discourse*. London: Routledge.

Teubert, W. 1996. Comparable or parallel corpora? *International Journal of Lexicography* 9(3): 238–264. https://doi.org/10.1093/ijl/9.3.238

Tognini-Bonelli E. 2000. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.

Vande Kopple, W. J. 1985. Exploratory discourse on metadiscourse. *College Composition and Communication* 36(1): 82–93. https://doi.org/10.2307/357609

Vassileva, I. 1998. Who am I/who are we in academic writing? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian. *International Journal of Applied Linguistics* 8(2): 163–190. https://doi.org/10.1111/j.1473-4192.1998.tb00128.x

Viaggio, S. 2006. *A General Theory of Interlingual Mediation*. Berlin: Frank &Timme.

Vinay, J-P. & Darbelnet, J. 1995[1958]. *Comparative Stylistics of French and English. A Methodology for Translation* [Benjamins Translation Library 11]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.11

Volansky, V., Ordan, N. & Wintner, S. 2013. On the features of translationese. *Literary and Linguistic Computing* 30(1): 98–118.

Wakabayashi, J. 2009. Translational Japanese: A transformative strangeness within. *PORTAL Journal of Multidisciplinary International Studies* 6(1). <http://epress.lib.uts.edu.au/journals/index.php/portal/issue/view/56 (11 August 2018). https://doi.org/10.5130/portal.v6i1.848

Xiao, Z. 2011. Word clusters and reformulation markers in Chinese and English: Implications for translation universal hypotheses. *Languages in Contrast* 11(2): 145–171. https://doi.org/10.1075/lic.11.2.01xia

Epistemic *must* in an English-Swedish contrastive perspective

Karin Aijmer
University of Gothenburg

Must in English and *måste* in Swedish do not mean quite the same thing and they have different formal and functional properties as can be shown from a translation perspective. They occur for example with different frequencies in the two languages. *Must* is mainly epistemic while the deontic meaning predominates with *måste*. Several factors play a role for the semantic-pragmatic function of *must* such as the degree of speaker certainty and the type and source of evidence. It is shown that *must* is used both to express the speaker's stance and with hearer-appeal in situations where the hearer is assumed to know more than the speaker. *Måste* is closely associated with certainty and with inferential meaning.

1. Introduction

In his writings Wolfgang Teubert has emphasised the use of parallel corpora as a resource for establishing translation equivalents between languages (Teubert 1996). However, translation equivalent is an evasive term. As Teubert points out, a word in the source language may correspond to a list of different terms in the target language. This is illustrated in this study presented in honour of Wolfgang Teubert, where the starting-point is that we do not find 100% equivalence even when we compare cognates.

English *must* and Swedish *måste* are cognates. However even "similar modal auxiliaries" may "differ in complex ways from language to language, making translation difficult" (Haugen 1976: 80, quoted from Palmer 1986: 35). The cognates *must* and *måste* are not always translation equivalents. This raises the question how similar they are and whether they have the same modal status. Cases where *must* does not correspond to *måste* (or vice versa) are of special interest because they may indicate areas where the inventory of forms used to express epistemic modality differs across languages (cf. Johansson 2007: 25).

https://doi.org/10.1075/scl.87.09aij

The aim of the present study is to discuss the similarities and differences between *must* and its Swedish cognate. The specific research questions are the following:

- What is the distribution and frequency of *must* and *måste* in English and Swedish original texts?
- When are *must* and *måste* translated by the cognate modal and when do translators choose a different translation?
- What do the translations by a non-cognate element show about the meanings of *must* and its cognates in other languages?

The paper is organised as follows. The material and method are discussed in Section 2. Section 3 defines epistemic modality and evidentiality and summarises the ongoing debate about the relation between them. Section 4 compares the frequencies of *must* and *måste*. Section 5 presents the Swedish correspondences of the English *must* and Section 6 discusses the modal and evidential values of Swedish *måste* on the basis of its translations. Section 7 contains the conclusion.

2. Material and method

It is notoriously difficult to distinguish between different meanings on the basis of intuition alone. If a lexical item 'a' in one language is rendered by a lexical item 'b' in the other language we can assume that 'a' and 'b' have the same meaning. An advantage with using translation equivalence as a criterion for sameness of meaning is that the context is the same in the two texts. Translators look for an appropriate matching word in the target language on the basis of analysing the linguistic and extra-linguistic context.

Studies in contrastive linguistics have increasingly made use of electronic corpora containing parallel texts in two or more languages. Parallel corpora are an important resource because of their size and the type of searches they allow. The present study is based on the texts in the *English-Swedish Parallel Corpus* (ESPC)[1] (Altenberg & Aijmer 2000). The corpus contains roughly 2.8 million words representing both fiction and non-fiction. I have used the fiction part of the corpus only since fiction and non-fiction can be expected to have different patterns of frequency and use. In Aijmer (2017) it is for example shown that *must* and *måste* were more frequent in non-fiction than in fiction and that they have different functions depending on the genre. In a bidirectional corpus such as the

1. <http://www.sprak.gu.se/forskning/korpuslingvistik/korpusar-vid-spl/> Description available at <http://sprak.gu.se/forskning/korpuslingvistik/korpusar-vid-spl/espc>

English-Swedish Parallel Corpus the starting-point for the searches can be either original texts or translations. Thus I first extracted all the occurrences of English *must* from the original fiction texts with their translations and then repeated the same procedure for Swedish *måste*.

3. Epistemic and evidential meaning

Must has both deontic (obligation) meaning and epistemic meaning. It also has some characteristics associated with evidentiality. The epistemic meaning is defined by van der Auwera and Plungian in terms of a judgement by the speaker: "a proposition is judged to be uncertain or probable in relation to some judgment" (van der Auwera & Plungian 1998: 81).

Evidentiality brings in a discussion of the type of evidence speakers have for their statements. *It is said that* can for example be understood as an evidential marker with the meaning "hearsay" (Cornillie 2007: 10). On the other hand, it does not refer to the speaker's epistemic assessment of how things are (Cornillie 2007: 10). The status of *must* is more controversial since there is no clear distinction between evidentiality and epistemic modality ('judgement of the truth of a proposition'). Palmer (1986: 70) suggests that the analysis of *must* is "a matter of evidentiality" and "that there is often no clear distinction between judgments [epistemic modality] and evidentiality since speakers' judgments are naturally often related to the evidence they have."

Van der Auwera and Plungian (1998), on the other hand, welcome a sub-type of evidentiality termed inferential which can be used to analyse both the Turkish evidential *-mis* and English *must*. This sub-type "identifies the evidence as based on evaluation or judgement (by the speaker) of whether a proposition is true" (van der Auwera & Plungian 1998: 85). Their claim is "that the inferential reading amounts to epistemic modality and more particularly epistemic necessity". Inferential modality is "therefore regarded as an overlap category between modality and evidentiality" (van der Auwera & Plungian 1998: 86).

Translation equivalence between elements in two languages is based on similarity of meaning. Both *must* and *måste* can be analysed as inferential-evidential. However, a closer analysis may show that there are differences between the auxiliaries which are reflected in how they are translated. If *must* is not translated into its cognate, the reason may for example be that languages analyse the relationship between epistemic modality and evidentiality differently.

4. Frequencies

Must and *måste* appeared with different frequencies in the original English and Swedish texts. There were 210 occurrences of *must* in the English original texts to be compared with more than twice as many occurrences in Swedish originals. However, many of the examples of *must* are deontic, that is they have meanings associated with obligation or (deontic) necessity. Table 1 gives the frequencies of *must* with different meanings (deontic and epistemic) in the *English-Swedish Parallel Corpus*.

Table 1. Differences in the frequency of *must* and *måste* on the basis of original and translated English and Swedish fiction texts in the ESPC

	English *must*		**Swedish *måste***	
	Eng original Swe translation	**Eng translation Swe original**	**Swe original Eng translation**	**Swe translation Eng original**
Epistemic	129 (61.4%)	164 (51.9%)	109 (24%)	108 (24%)
Deontic	77 (38.6%)	158 (49.1%)	345 (76%)	343 (76%)
Total	210 (100%)	322 (100%)	454 (100%)	450 (100%)

The table shows that English *must* is primarily epistemic while Swedish *måste* is most often deontic. The epistemic meanings in the Swedish (original) texts make up only 24% of the examples of *måste* to be compared with 61.4% of *must* in the English originals. The high frequency of epistemic meanings of *must* has also been noticed by Biber et al. (1999: 494): "surprisingly, *must* in conversation is used most of the time to mark logical necessity" (epistemic meaning). These differences in frequency are interesting in a wider cross-linguistic perspective. On the basis of comparing English *must* and its translations into German and Dutch, Mortelmans (2010, 2012) found for example that "despite the fact that the English modal verb *must* is a highly grammaticalised modal, it has a considerably smaller frequency of occurrence than its immediate German and Dutch counterparts *müssen* and *moeten*" (Mortelmans 2012: 2150).

5. The Swedish correspondences of English *must*

The cognates *must* and *måste* are not always translation equivalents. This raises the question how similar they are and what the alternatives are when they are not chosen as each other's correspondences. Table 2 shows the correspondences of the English *must* on the basis of the occurrences in originals and sources in the ESPC.

Table 2. The Swedish correspondences of English *must* in the ESPC

	Eng original Swe translation	Eng translation Swe original	Total
måste	104 (80.6%)	98 (68.9%)	202 (70.6%)
väl [I suppose]	6	19	25
måtte [must]	5	3	8
säkert [certainly, surely]	3	4	7
ju [as you know]	5	1	6
nog [probably]	–	6	6
kan [can]	2	1	3
alltså [consequently]	–	3	3
borde [should, ought to]	–	3	3
förstås [of course]	–	3	3
skulle [would]	1	1	2
visst [certainly, someone says that]	–	2	2
antagligen [probably]	1	–	1
kunde [could]	–	1	1
tydligen [obviously]	–	1	1
förmodligen [probably]	–	1	1
troligen probably]	–	1	1
ändå [anyhow]	–	1	1
ø	1	13	14
Paraphrase	1	2	3
Total	129	164	293

Note: In examples where the modal auxiliary is modified by a particle, the auxiliary and the particle have been given separate entries in the table.

According to Smith (2003: 262), "within the fiction section close to half of the occurrences [of epistemic *must*] are in dialogue".[2] This was confirmed in my study where about 47% of the examples (61/129) were found in dialogue context (direct speech).[3] By far the most frequent correspondence in both directions is the cognate *måste*. A striking fact is that in translations from English into Swedish the frequency of *måste* is even higher than in the translation of Swedish source texts into English. A plausible explanation for this imbalance is that the similarity between *must* and *måste*

2. Smith's study was based on the LOB and the FLOB corpora.

3. However only 25% of the occurrences were found in direct speech in the Swedish sources (42/164).

makes the "direct" translation easier in the direction from English into Swedish. The frequencies of *must* in translations from Swedish and in Swedish sources have therefore been used to calculate the degree to which they (mutually) correspond. As shown in Table 2, *must* and *måste* are translated into each other in 70% of the cases (on intertranslatability or mutual correspondence, see Altenberg (1999)).

Swedish has many different ways of expressing epistemic or evidential meaning in addition to the cognate modal auxiliary. The most frequent correspondence in both Swedish translations of English *must* and Swedish source texts with English translations was the particle *väl* (see Section 5.2). However, even translation alternatives which are of low frequency can point to new or emergent pragmatic functions.

In a number of cases, translations of *must* are rendered with lexical items belonging to different word classes. The non-congruent correspondences (*must* not = *måste*) include modal adverbs and modal particles as well as other modal auxiliaries. Zero omissions can be regarded as a special category (see Table 3).

Table 3. Swedish correspondences of English *must*: type of category

	Eng original Swe translation	**Eng translation Swe original**
Modal auxiliary	112	107
Modal particle	11	28
Modal adverb (epistemic or evidential)	4	14
Zero	1	13
Paraphrase	1	2
Total	129	164

Modal particles (or combinations with modal auxiliaries and modal particles) were the most frequent way of expressing epistemic modality in the translation correspondences (after a modal auxiliary). The adverbial correspondences are adverbs associated with epistemic qualification (*säkert, antagligen, förmodligen, troligen)* or evidential adverbs (*tydligen, förstås, alltså, ändå)*.

5.1 *Must* translated by expressions of certainty

When *must* corresponds to (epistemic) adverbs it expresses certainty rather than possibility. In (1) *must* is translated as *säkert* [certainly] and in (2) as *antagligen* [probably] (see Examples 1 and 2):[4]

4. For detailed information about the text codes in brackets after examples, see Altenberg and Aijmer (2000) and Altenberg, Aijmer and Svensson (2001).

(1) It must be something to do with my relationship with my father." (FW1)
Säkert har det med mitt förhållande till pappa att göra." (FW1T)

(2) Such a beautiful deep blue they'd made it, when it must have been almost colourless to begin with. (DF1)
De hade lyckats få en sån fin blå färg på dem, och de hade antagligen varit nästan färglösa från början. (DF1T)

According to Nuyts (2001: 228)

> the question remains what differentiates the modal and the adverb in a speaker's choice of an epistemic expression. The only real semantic difference we have observed is the precision of the epistemic qualification expressed by each: while the adverbs refer to specific positions on the epistemic scale, the modal auxiliaries are much vaguer.

The translator's need to be specific may explain the translation as *säkert* [certainly] or a "weaker" epistemic adverb (*förmodligen* [probably], *antagligen* [probably], *troligen* [probably]).

5.2 *Must* translated by epistemic particles

As shown in Table 3, *must* was often rendered by a modal particle in Swedish. Modal particles are little words such as *väl* [I suppose], *nog* [probably], *ju* [as you know], *visst* [certainly, someone says that]. The particles can be analysed as epistemic and evidential and they can be oriented to both the speaker and the hearer (and to "what other people say").

Taken together, the occurrences with *väl* (with the paraphrases 'I suppose', 'I guess', 'I think') constitute one of the most frequent correspondence patterns. It was represented in 25 examples (either in the original text or in the translation) and was therefore the most frequent translation choice. However, *väl* was more characteristic as a Swedish source item (*väl* > *must*) than as a translation reflecting the fact that translators take cognate modal auxiliaries to be the best equivalent even though the particle might be more appropriate. *Väl* is epistemic rather than evidential. It expresses the speaker's uncertainty and is oriented to the hearer as the source of knowledge: "by means of *väl* the speaker indicates that the listener has access to other knowledge of the state of affairs than what he gets from the speaker" (translated from Teleman, Hellberg & Andersson 1999: 116). Like other epistemic expressions it can be used strategically to achieve certain effects (cf. Nuyts 2001: 44). *Väl* can for instance be used for reasons of politeness to express the speaker's weak involvement rather than uncertainty. According to Mortelmans, the speaker

> does not primarily want to indicate a high degree of probability of the state of affairs or evaluate the evidential bases for a particular claim, but rather uses epistemic *must* as a means to appeal to the addressee, to invoke his or her sympathy, to contribute to the addressee's positive face. (Mortelmans 2012: 2159)

In Example (3), *you must miss working on the land* suggests that the speaker sympathises with how the hearer must be feeling although the speaker has no knowledge about whether this is the case or not:

(3) "I was in the country a lot when I was small," the stranger said.
"How wonderful it was with all that work. Think how everyone did their job, and with such joy. You must miss working on the land." (SC1T)
Jag var mycket på landet då jag var liten, sa främlingen.
Tänk vad underbart det var med allt arbete. Tänk hur alla jobbade och med vilken glädje. Ni saknar väl arbetet med jorden. (SC1)

In Example (4) the lady uses *must* [affably] to express interest in the hearer's doings:

(4) Då var väl televisionen en upplevelse för er, sa damen vänligt. (SC1)
"Television must have been quite an experience for you," said the lady affably. (SC1T)

Example (5) illustrates that "intersubjective epistemic *must* is often found with predicates that denote a somehow negative state of affairs for the hearer, with which the speaker empathizes" (Mortelmans 2012: 2160). *I guess* (Swedish *förstår jag*) seems to express more certainty that something is the case and increases the speaker's involvement:

(5) I guess you all must be hungry. (GN1)
Ni är väl hungriga båda två, förstår jag. (GN1T)

In (6) *must* is also used with an intersubjective and hedging function. *Ju* in the translation is hearer-oriented like *väl* but presupposes that the hearer knows ('as you know', 'as you should know') (Teleman, Hellberg & Andersson 1999: 114):[5]

(6) "Marion Hopfoot," said Hilary.
"She's his secretary. …
Well, you must have known. (FW1)
"Marion Hopfoot", sa Hilary.
"Det är hans sekreterare. …
Fast det måste ni ju ha vetat om. (FW1T)

5. By means of *ju* the speaker indicates that s/he takes into account (or pretends) that both the speaker and the hearer (or people in general) know or can deduce from what has been said in the sentence that something is true: "as you and I know" (translation from Teleman, Hellberg & Andersson 1999: 114).

Must is not primarily used to express a high degree of certainty or probability. The translator has added *ju* which conveys that *must* has a deontic overlay of meaning ('you should have known this').

Besides *väl* there were 6 examples of *nog* in Swedish sources (a particle signalling high probability and paraphrasable as 'I suppose' (Teleman, Hellberg & Andersson 1999: 117)). The speaker does not have evidence for claiming that something is the case but makes a conjecture. Notice the co-occurrence with *tänker jag/I think* marking the speaker-based source for the claim in Example (7):

(7) Han är nog omkring femtio – fast han ser yngre ut, tänker jag och ser in i hans blick. (MS1)
He must be around fifty but looks younger, I think, and look into his eyes. (MS1T)

In Example (8) *must* has been used to translate *visst* (*visst* [from what I have heard, as it seems], Teleman, Hellberg & Andersson 1999: 109):

(8) På hemvägen frågade hon Karin vem det var som hälsat på hos mormor. Karin tittade länge på vägen innan hon svarade. Hon var inte riktigt säker, sa hon, det var visst nån släkting till mormor. (MG1)
On the way home she asked Karin who the visitor was. Karin looked at the road a long time before she answered. "I'm not exactly sure," she said. "She must be one of Grandmother's relatives. (MG1T)

Mortelmans (2012: 2152) did not find any examples where English *must* has the meaning hearsay which suggests that this meaning is infrequent. On the other hand, Dutch *moeten* has developed a reportative reading in addition to the inferential meaning it shares with *must* (De Haan 2009).

5.3 *Must* with inferential meaning

Translations of *must* with *alltså, förstås, följaktligen, tydligen* suggest that *must* should be interpreted as inferential. In Example (9) the speaker has just realised that the addressee "is not his brother but his brother-in-law". From this he can conclude that "you are married to his sister":

(9) "Are you his brother?"
"I'm his brother-in-law, Eugene Nickerson."
"You must be married to his sister then," I said. (SG1)
"Är ni bror till honom?" frågade jag på försök.
"Jag är hans svåger. Mitt namn är Eugene Nickerson."
"Då är ni alltså gift med hans syster", sade jag. (SG1T)

Both *då* and *alltså* convey the inferential meaning of *must*. The speaker draws a conclusion on the basis of reasoning from evidence.

In Example (10) *must* co-occurs with *surely* in the original to express a combination of inferential meaning with deontic meaning ("the bank should be interested in what happened to the money"):

(10) Surely, the bank must care what happened to twenty-five thousand dollars. (SG1)
Banken borde väl ändå vara intresserad av vad som hände med tjugofemtusen dollar. (SG1T)

5.4 Zero-correspondence

Zero-correspondence is an indication that "the resources in languages do not match, or it reflects differences with respect to the relevance of marking a particular type of meaning" (Johansson 2007: 94). The number of examples of zero-omission in the translation direction from English originals to Swedish translations is low but the examples of addition in the other translation direction is the most frequent correspondence after the particle *väl*. In Example (11) *must* has been omitted in the translation into Swedish:

(11) People who meet her for the first time at a social gathering tend to treat her politely as extrovert Ted's quiet and dependable wife, and imagine that she must be a great support to him. (MD1)
När folk träffar henne första gången i sällskapslivet ser de henne gärna som den extroverte Teds lugna och pålitliga hustru och föreställer sig att hon är ett stort stöd åt honom. (MD1T)

The meaning of *must* is sufficiently expressed by the higher clause ('people imagine that') and has not been rendered by the translation.

However, in Example (12), *must* in the English translation seems to have appeared "out of the blue" (Johansson 2007: 26):

(12) Men hur mycket de behövde sin dagliga resa förstod jag först när jag fick veta att de bodde i Köpenhamn och var tvungna att stiga upp halv fyra på morgonen för att ta bussen till Dragör. (BL1)
But I realised how much they must have needed it when I learned that they lived in Copenhagen and had to get up before four in the morning to catch the bus to Dragor. (BL1T)

There is no *måste* in the original Swedish text which can explain why the translator has chosen *must*. *Must* has been added to mark the speaker's subjective involvement with the people who needed to get up very early to take the bus.

In Example (13) the Swedish original contains the past tense. English inserts a redundant *must* to soften the negative effect of the assertion (*it must be serious*):

(13) – Jag är hemma. Kom hit.
Tydligen insåg Martinson att det var viktigt. (HM2)
"I'm at home. Get out here."
Martinson evidently realized it must be serious. (HM2T)

To sum up, when *must* is not rendered as *måste* which represents the easiest translation, it could be because *must* has extended its meaning to new pragmatic functions. The source of the evidence is a particularly important aspect of *must*'s use as shown by its translations. It appears that the inferential meaning can be weakened and that *must* is used both with speaker-oriented epistemic meaning and when the source of evidence is the hearer (the hearer-oriented or intersubjective meaning). When the epistemic-inferential *must* is speaker-oriented it expresses probability and can be paraphrased as 'I suppose'. It can also be used to appeal to what the hearer knows (the speaker assumes that the hearer has better knowledge about the state of affairs than she has herself). The hearer-oriented or intersubjective use of *must* has developed a number of functions associated with the speaker's personal involvement rather than with probability. The hearer-oriented *must* can also convey a deontic meaning (associated with what should or ought to be the case) as in Example (10) above.

Must was sometimes omitted or added in the translation. The ease with which it can be inserted in English although it does not have a formal correspondent in the source language suggests that (the intersubjective) *must* is a characteristic feature of English conversation. The results from the present study can be compared with the use of other cognates of *must*. In German (and Dutch) Mortelmans (2012) for example found a similar tendency not to translate epistemic *must* with the related *müssen* or *moeten* when *must* has intersubjective function.

6. The English correspondences of Swedish *måste*

In the Swedish Academy Grammar (Teleman, Hellberg & Andersson 1999: 308), (the epistemic) *måste* is defined as follows:

> with epistemic meaning the verb *måste* …indicates in its basic meaning that the sentence is necessarily true on the basis of what the speaker has concluded by taking into account the available facts…. The use of *måste* does not imply that the speaker guarantees the truth of the sentence only that all the available data indicate that what is said is true. (author's translation)

In the sentence below (Example 14) (from Teleman, Hellberg & Andersson 1999) the speaker is not certain but the evidence allows him/her to conclude that "I must have left the keys at home":

(14) Jag måste ha glömt nycklarna hemma
[I must have left the keys at home]

Table 4 shows the correspondences of the Swedish *måste* in translations into English and in the Swedish sources.

Table 4. The English correspondences of Swedish *måste* in the ESPC

	Swe original Eng translation	Swe translation Eng original	Total
must	91	95	186
surely	1	3	4
would	2	2	4
have (got) to	3	–	3
could	2	–	2
of course	1	1	2
will	–	1	1
can	1	–	1
probably	–	1	1
sounds like	1	–	1
may	–	1	1
it is certain to	–	1	1
zero	5	2	7
paraphrase	2	1	3
Total	109	108	217

The mutual correspondence between *måste* and *must* is high. The Swedish modal auxiliary is predominantly translated by its English cognate and the English source is *must* in most cases. *Måste* has fewer translation alternatives than *must* and apart from *surely* (*would* and zero) these are represented only once or twice (cf. Table 4).

When translated by a cognate *måste* is inferential. The speaker draws a conclusion on the basis of direct or indirect evidence (Example 15):

(15) Du är tjugotre år. Det måste vara femton år mellan er.
Låt honom vara. Han är farlig. (KE1)
You're twenty-three. There must be fifteen years between you. Keep away from him. He's dangerous. (KE1T)

In Example (16) the simple *måste* is avoided in the translation of *surely. Måste väl* expresses the hearer-appealing function of the original: "(surely) it must have occurred to you that the message is directed at you".

(16) And why the flowers, for God's sake? What were they supposed to tell us?"
"You'll have to talk to the police," said Robin with sudden decision, reaching for the telephone.
"Dammit, Sarah, who else knew she called you her scold's bridle? Surely it's occurred to you that the message is directed at you." (MW1)
Och vad skulle blommorna vara bra för? Hur var det tänkt att vi skulle tolka dem?"
"Du måste prata med polisen", sade Robin med plötslig beslutsamhet och sträckte sig efter telefonen.
"För fan, Sarah, vem mer visste att hon kallade dig sitt trätobetsel? Det måste väl ha slagit dig att budskapet kan vara riktat till dig?" (MW1T)

In Example (17) *måste nog* conveys the speaker's hesitation and uncertainty:

(17) Ja, det måste nog vara jag, sa Martin Beck tveksamt. (SW1)
"Yes, I guess that must be me," said Martin Beck uncertainly. (SW1T)

The Swedish original expresses epistemic modality "more narrowly" than in English where the same modal meaning is expressed in both the matrix clause and the subordinate clause.

6.1 *Måste* translated by expressions of certainty

Translations with an expression of certainty (*surely, I'm sure, it's certain)* seem to strengthen the evidential meaning, see Example (18):

(18) Till och med min paratyfus stämde inte med de andra barnens. För säkerhets skull fick jag en spruta i varje lår. Den gången sprallade och skrek jag så mycket att de fick vara tre som höll mig och en som gav sprutorna. Sådant måste stå i journaler. (RJ1)
Not even my paratyphoid was like other children's. Just to be on the safe side, they gave me an injection in both thighs. I wriggled and screamed so much that they needed three people to hold me down and a fourth to put the needle in. I am sure that that sort of thing ends up in your medical record. (RJ1T)

It's certain to (Example 19) in the English source text is used rhetorically to strengthen the argument ('if the book is written by an American it must be filth'):

(19) "If it's by an American it's certain to be filth. (RD1)
"Om det är en amerikansk författare som har skrivit den, så måste det vara skräp. (RD1T)

In Example (20) *måste* has been translated by the combination *must surely. Surely* is used to emphasise the speaker's certainty:

(20) Med hans kunskaper om hav måste han ha haft något med sjön att göra. (BL1)
As he seemed to know so much about sailing, he must surely have something to do with the sea and ships. (BL1T)

6.2 *Måste* translated by expressions of evidentiality

Among the translation alternatives are also *will* and *would*. Both are closely related to *must* in the inferential reading. According to Palmer (2003: 8), "epistemic *MUST* and *WILL* have some characteristics of Evidential modality, for they signal conclusions that are based on evidence". However, the evidential *will* refers to a situation in the future ('you can be sure that everything will be found in the papers'):

(21) "She kept her papers very neat," said Mrs Spede eventually, after receiving some sort of permission to speak from her husband.
"It'll all be in the papers." (MW1)
"Hon har väldig ordning i sina papper", sade fru Spede till slut sedan hennes make givit henne något slags tillstånd att tala.
"Allting måste finnas bland papperen." (MW1T)

Would co-occurs with *likely* rather than with *surely* (Example 22):

(22) Här var en trädörr och den måste väl rimligtvis leda in till en bastu. (LG1)
There was a wooden door here and that would quite likely lead into a sauna. (LG1T)

Summing up, the high number of congruent translations (*måste* = *must*) and the absence of alternatives with epistemic meaning suggests that Swedish *måste* is more inferential and less epistemic than the English cognate. *Måste* was only used with intersubjective meanings when it had epistemic support from a particle in Swedish.

7. Conclusion

Must and *måste* are not always each other's translation equivalents. When the translator chooses a less direct alternative this may indicate that the languages differ and that they structure the field of evidential-epistemic modality differently. In other words, *must* in English and *måste* in Swedish do not mean quite the same thing and they have different formal and functional properties. They occur for example with different frequencies in the two languages.

Several semantic factors play a role for the semantic-pragmatic function of *must*, such as the degree of speaker certainty associated with the speaker's epistemic commitment to what is said, the type of evidence (for example whether the speaker bases herself on indirect (e.g. perceptual) evidence, on guesses and assumptions or on hearsay evidence, the source of evidence (the speaker, hearer, or 'other' as the one who knows best). *Must* is for example used both to express the speaker's subjective stance and with 'hearer-appeal' in situations where the speaker assumes that the hearer knows more than the speaker.

On a deeper level, the differences between *must* and its cognates involve grammaticalisation and semantic changes such as subjectification and intersubjectification. The extent to which modal auxiliaries are associated with deontic or epistemic modality correlates with their degree of grammaticalisation. The semantic processes apply in a uniform way across languages. As a part of this process the strongly inferential meaning becomes weakened and replaced by "a more strongly subjective-epistemic component" (Mortelmans 2000: 133). The epistemic meanings of *must* can for example be regarded as more subjective ("increasingly based in the speaker's subjective belief/attitude toward the proposition" (Traugott & Dasher 2002: 95)).

Måste is more frequent with deontic than with epistemic meaning indicating that it is less grammaticalised than English *must*. The inferential meaning predominates as shown by the translations with a cognate. *Måste* is closely associated with certainty (the speaker's strong confidence) rather than with uncertainty or hearer-appeal. However, combinations of *måste* with other modal elements with epistemic meaning (in translations and in originals) suggest that *måste* can be extended to new uses in the interaction.

It is important to compare related lexical items in many different languages to confirm tendencies based on two languages only. *Must* and its cognates in German and Dutch have also been analysed contrastively in several studies by Mortelmans. The present study suggests that Swedish *måste* shares more features with its German and Dutch cognate than with *must*. *Måste* and its cognates in German and Dutch are for example less frequent than *must* in the epistemic function and are less typically used in dialogic or interactive contexts with more interpersonal meanings.

References

Aijmer, K. 2017. The semantic field of obligation in an English-Swedish perspective. In *Contrastive Analysis of Discourse-pragmatic Aspects of Linguistic Genres* [A special issue of the *Yearbook of Corpus Linguistics and Pragmatics*], K. Aijmer & D. Lewis (eds), 1–32. New York NY: Springer. https://doi.org/10.1007/978-3-319-54556-1_2

Altenberg, B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In *Out of Corpora. Studies in Honour of Stig Johansson*, H. Hasselgård & S. Oksefjell (eds), 249–268. Amsterdam: Rodopi.

Altenberg, B. & Aijmer, K. 2000. The English-Swedish parallel corpus: A resource for contrastive research and translation studies. In *Corpus linguistics and linguistic theory. Papers from the 20th International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999*, C. Mair & M. Hundt (eds), 15–33. Amsterdam: Rodopi.

Altenberg, B., Aijmer, K. & Svensson, M. 2001. *The English-Swedish Parallel Corpus (ESPC): Manual*. Department of English, Lund University. <http://www.sol.lu.se/engelska/corpus/corpus/espc.html> (15 January 2017).

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Cornillie, B. 2007. *Evidentiality and Epistemic Modality in Spanish (Semi-)auxiliaries. A Cognitive-functional Approach*. Berlin: Mouton de Gruyter.

de Haan, F. 2009. On the status of 'epistemic' *must*. In *Modality in English 3*, R. Facchinetti & A. Tsangalidis (eds), 261–284. Bern: Peter Lang.

Haugen, E. 1976. *The Scandinavian Languages: An Introduction to Their History*. London: Faber and Faber.

Johansson, S. 2007. *Seeing Through Multilingual Corpora. On the Use of Corpora in Contrastive Studies* [Studies in Corpus Linguistics 26]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.26

Mortelmans, T. 2000. On the 'evidential' nature of the 'epistemic' use of the German modals *müssen* and sollen. *Belgian Journal of Linguistics* 14: 131–148.

Mortelmans, T. 2010. Falsche Freunde: Warum sich die Modalverben *must, müssen* und *moeten* nicht entsprechen. In *Modalität/Temporalität in kontrastiver und typologischer Sicht*, A. Katny & A. Socka (eds), 133–148. Frankfurt: Peter Lang.

Mortelmans, T. 2012. Epistemic *must* and its cognates in German and Dutch. The subtle differences. *Journal of Pragmatics* 44: 2150–2164. https://doi.org/10.1016/j.pragma.2012.09.011

Nuyts, J. 2001. *Epistemic Modality, Language, and Conceptualization* [Human Cognitive Processing 5]. Amsterdam: John Benjamins. https://doi.org/10.1075/hcp.5

Palmer, F. R. 1986. *Mood and Modality*. Cambridge: CUP.

Palmer, F. R. 2003. Modality in English: Theoretical, descriptive and typological issues. In *Modality in Contemporary English*, R. Facchinetti, F. R. Palmer & M. Krug (eds), 1–18. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110895339.1

Smith, N. 2003. Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In *Modality in Contemporary English*, R. Facchinetti, M. Krug & F. R. Palmer (eds), 241–266. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110895339.241

Teleman, U., Hellberg, S. & Andersson, E. 1999. *Svenska Akademiens grammatik*, Band 4. Stockholm: Norstedt.

Teubert, W. 1996. Comparable or parallel corpora. *International Journal of Lexicography* 9: 238–264. https://doi.org/10.1093/ijl/9.3.238

Traugott, E. C. & Dasher, R. 2002. *Regularity in Semantic Change*. Cambridge: CUP.

van der Auwera, J. & Plungian, V. 1998. Modality's semantic map. *Linguistic Typology* 2: 79–124. https://doi.org/10.1515/lity.1998.2.1.79

Translating fictional characters – *Alice* and *the Queen* from the Wonderland in English and Czech

Anna Čermáková and Michaela Mahlberg
University of Birmingham

In this chapter, we propose a novel theoretical framework for the literary translation of fictional characters. This framework develops the cognitive corpus linguistic notion of mind-modelling to account for process-, product- and function-oriented aspects of literary translation. We use the examples of Alice and the Queen from *Alice's Adventures in Wonderland* to compare character cues across the English original and a Czech translation. The character cues we focus on are reporting verbs. Reporting verbs, as part of the presentation of fictional speech, form a central component of narrative fiction and so provide an ideal evidential basis for our theoretical framework. The translation shifts we found through our comparison of source and target text specifically include gendered uses of reporting verbs. By approaching the target text as both a translation and a reading of the text in its own right we are able to view translation shifts as a reflection of shifts in the mind-modelling of fictional characters.

1. Introduction

Translation can be described as an actitvity whereby "[a] translator produces a paraphrase of a text in another language" (Teubert 2002: 190–191). This description sounds straightforward. However, the complexity of accounting for features of translated literary texts is reflected by the combination of different disciplines that are relevant to the undertaking. Translation studies in itself is characterised by a plurality of approaches that can concentrate on the descriptive or applied aspects of the discipline and distinctions can be made between translation as product-, process- or function-oriented. Literary translation specifically can be approached under the heading of specialised translation or viewed in terms of the history of the discipline where the translation of literary texts plays a fundamental role. Wittman (2013: 438) defines 'literary translation' as "the product of a translator

https://doi.org/10.1075/scl.87.10cer

who takes seriously the literary nature of the original and translates with the goal of producing a text that will have literary merit of its own". In this chapter, our focus is on the translation of fictional characters. The creation of fictional characters contributes to the literariness of a text and hence deserves particular attention from a translation point of view. The linguistic techniques of characterisation that are employed by an author are of interest to stylisticians as well as literary scholars, the latter placing particular emphasis on how characters can be situated within the social and cultural contexts of the time of writing.

The text we focus on in this chapter is Lewis Carroll's *Alice's Adventures in Wonderland* (1865). The *Alice* books (the subsequent *Through the Looking Glass* was published six years later in 1871) firmly belong to the canon of children's literature all over the world. Although published 150 years ago, they still have an avid readership not to mention their attractiveness to the film industry. Research into the *Alice* phenomenon extends well beyond literary studies.[1] When it was first published, *Alice's Adventures in Wonderland* signalled a change in the writing style for children (Carpenter 2009). The *Alice* books marked the beginning of 'fantasy' as a literary genre in children's literature but it is the language of 'nonsense' that sets the books apart from the rest of their contemporaries. Both fantasy and nonsense subvert traditional narrative rules and expectations regarding the genre of children's books. In the Wonderland, nonsense does not function solely on the word level but also affects conversation and the linguistic conventions of communication. Conversation becomes a game, which is increasingly difficult to sustain as basic pragmatic principles do not hold any longer.

The *Alice* books might be best known for their approach to nonsense and the deviations from conventional communicative patterns. However, in this chapter, we focus on seemingly less striking linguistic elements that nevertheless can function as textual cues for the creation of fictional characters. We will approach the translation of fictional characters within a cognitive corpus linguistic framework that conceptualises characterisation through mind-modelling (Stockwell 2009). Mind-modelling refers to the fundamental human capacity "for imagining and maintaining a working model of the characteristics, outlook, beliefs, motivations and consequent behaviour of others" (Stockwell & Mahlberg 2015: 132). In the context of literary texts, it accounts for the readerly process of modelling the minds of fictional characters. Mind-modelling assumes that this process begins with a template based on the reader's own sense of person-ness. Important to a corpus linguistic approach to mind-modelling is that textual cues provide character

1. An example is the exhibition "The Alice Look" at the Victoria & Albert Museum of Childhood curated by Kiera Vaclavik in 2015, the exhibition explored Alice as a fashion icon, see <http://www.vam.ac.uk/moc/exhibitions/the-alice-look/> (15 January 2018).

information that differentiates the template away from the self of the reader towards the representation of the fictional character. Corpus linguistic methods support the identification of such textual cues and patterns in the literary text.

In this chapter, our outline of the cognitive corpus linguistic framework of mind-modelling is a very condensed summary of the theory outlined in more detail elsewhere, see specificially Stockwell and Mahlberg (2015) and Mahlberg and Stockwell (2016). Crucial for our approach to the translation of fictional characters is that mind-modelling is a text-driven process where information in the text forms an evidential basis for the creation of a reader's mental model of characters. As our approach to mind-modelling is a corpus linguistic one, we are specifically concerned with repeated phenomena or patterns in the text. This is not to say that one-off pieces of information are not relevant. But the focus on patterns acknowledges that there is overlap between fictional and general language patterns – which is a reflection of the continuities between fictional and real worlds. These continuities enable us to account for the plasticity of character cues in translation.

While different areas of translation studies focus on translation as a process, consider the product of translation or its function, looking at literary translation through the lens of mind-modelling highlights how all three are connected. As a reader, the translator begins with a template based on their self-hood to arrive at a representation of a fictional character and at the same time the translator has to translate character cues that will shape the reader's (i.e. the reader of the translation) template in the creation of a fictional character. Both for the translator and for the reader of the translated text the process of mind-modelling will be text-driven. However, like readers in general, translators are not necessarily aware of all the textual cues that shape the way in which they arrive at mental representations of fictional characters. This will be an important point for a corpus linguistic approach to literary translation.

A key contribution of corpus linguistics to the study of language more widely is an increasing range of methods that can help to identify textual patterns that language users might not easily be able to notice. In the field of translation studies, such methods have helped to reveal types of differences between translated and non-translated language (e.g. Mauranen & Kujamäki 2004, Laviosa 2002, Xiao 2010, Kruger, Wallmach & Murray 2011). Translation studies is one of the many areas of language study that corpus linguistics has had an impact on. Literary stylistics, which is concerned with the linguistic study of literature, is another such area. The term 'corpus stylistics' is now often used to refer to studies where corpus methods help to quantify, identify, systematise and compare patterns in fiction and literary texts more broadly (e.g. Semino & Short 2004, McIntyre 2010, Siepmann 2015, Shepherd & Berber Sardinha 2013, Louw & Milojkovic 2016). Together with

the development of corpus stylistics, there is emerging interest in corpus stylistic approaches to literary translation. The strengths of such approaches lies in the detailed comparison of source and target text to identify literary relevant shifts (e.g. Johnson 2016, Mastropierro 2017, Ruano San Segundo 2017, Toolan 2018, Čermáková 2018). Corpus methods can also provide strategies to support the translator in the process of translating (e.g. Čermáková 2015). In this sense, corpus approaches to literary translation focus on a specific text and literary features, but are similar to corpus linguistic approaches to translation studies more widely. In her introductory textbook, Olohan (2004) outlines the value of corpus methods to identify features of translation by studying both comparable and parallel corpora. She also highlights the potential of corpus linguistics in translator training and translation practice, as well as to describe and compare the styles of individual translators. Corpus methods can help describe stylistic features that a translator might not be aware of and so might reveal unconscious habits that affect the translation process. While corpus methods focus on translation as product, they can be combined with methods that provide insights into the cognitive processes of the translation process. For instance, Serbina et al. (2017) use keystroke logging and eye-tracking data to draw conclusions about cognitive processes that result in specific features of translations.

While corpus methods have brought new approaches to translations studies, the study of translated texts can in turn bring new dimensions to corpus linguistics and specifically to the study of literary texts. Potential that seems to be unexplored is to look at a translation as evidence of a reading of a literary text. When stylisticians analyse a text, they are concerned with the effects that linguistic features will have on the reader. The assessment of such effects tends to be made by the researcher as one reader. However, there is also a growing body of empirical work that complements the analysis of textual features with data on reactions and behaviours of readers – Whiteley and Canning (2017) provide a recent overview of such reader response research. Corpus methods are used to study textual features, so do not allow direct observations of readerly behaviour. But through comparisons of different texts and corpora, corpus methods can generate data that allow assumptions about effects of textual features. A common approach in corpus stylistics is to compare a specific text against a wider reference corpus to identity features that are characteristic of the individual text. The reference corpus then serves as a norm to describe the language experience that readers might bring to a text. So unusual textual features can be seen as potentially creating specific effects. This point has been strikingly made by Louw (1993) who uses the concept of semantic prosodies to account for such readerly effects. Corpus methods can also help to identify features in literary texts that are then linked to reading behaviour, as illustrated in an eye-tracking study by Mahlberg, Conklin and Bisson (2014) – a

methodological approach similar to Serbina et al. (2017) but focusing on literariness rather than translation.

As the potential of corpus linguistics to a range of fields of language study is most readily seen in its methodological offering, opportunities for frameworks that span disciplines have received little attention. With our approach to the translation of fictional characters we are able to extend the framework of mind-modelling beyond a monolingual conceptualisation. At the same time, we show that viewing a text as a product of the translation process is too narrow. A literary translation incorporates a reading of the source text as a literary text. The specific nature of children's literature and its translation (see e.g. Lathey 2015, 2011) adds an extra layer of complexity. Children's literature is written for children, but generally written and translated by adults. In terms of mind-modelling, the translator does not only work with their template of person-ness but also considers a child's version of this template. Overall, moving beyond an account of observable shifts in linguistic features, a cognitive corpus stylistic approach to the comparison of source and target texts enables a conceptualisation of translated fictional characters. In this chapter, we approach this theoretical framework with a textual analysis of two specific characters: *Alice* and *the Queen* in *Alice's Adventures in Wonderland* and in a Czech translation of the text.

2. Repetition and reporting verbs

The *Alice* books have been translated into more than 50 languages and in many languages they exist in more than one version, which makes them one of the most translated literary works (Horton 2002, Lindseth & Tannenbaum 2015). The *Alice* texts are challenging from the translator's perspective – especially the verses, puns, nonsense and invented words that only rarely translate "directly" (e.g. Weaver 1964, Kibbee 2003, Ambrosiani 2012). However, what we are interested in in this chapter are not unusual or striking words and phrases, but more subtle textual cues. As we have outlined above, it is a strength of corpus linguistic methods that they can help to find linguistic features that readers and translator are not immediately aware of. Corpus research has shown that repeatedly occurring patterns or highly frequent words tend to be among the subtle features of the language (e.g. Sinclair 1991, Mahlberg 2005). Crucial to the approach of mind-modelling that we follow in this chapter is the fact that corpus linguistic methods help to identify recurrent patterns that can contribute textual cues for the creation of fictional characters. Importantly, repeatedly occurring patterns can be more or less subtle. Whether patterns in a text appear to be striking or unobtrusive is linked to the frequencies with which such patterns appear more generally in the language. When a low

frequency word like *cockroach* is repeated several times within a few lines, a reader is not likely to miss this. Whereas the repeated occurrence of *the* or other function words within a short stretch of text will not appear as striking (unless the usage deviates in other ways from the norms of the language). Both striking repetitions and unobtrusively frequent words and patterns provide relevant textual cues for the creation of fictional characters. The latter, however, have received less attention by stylisticians or literary critics, as we have shown, for instance, in our earlier work on Dickens (cf. Mahlberg 2013).

While repetition as a stylistic device is relevant to the analysis of literary texts (cf. e.g. Toolan 2012, 2016; Leech & Short 2007), translators seem to view repetition as stylistically undesirable. The avoidance of repetition universally belongs to translation practice that can even operate subconsciously (Toury 1977, 1995; Ben-Ari 1998). In our previous work, we considered repetition in translation, for instance, to show how translators introduced shifts in cohesive networks (Mastropierro & Mahlberg 2017) or to consider the strategies that can be employed to support the translation process by considering repeated patterns first (Čermáková 2015). In fiction, reporting verbs are a particular type of repetition. Their occurrence is linked to the presentation of the speech of fictional characters. Character speech makes up a substantial part of most fictional texts. What fictional characters say, however, is more diverse than the range of verbs to report their speech, which is reflected in the repetition and frequency of reporting verbs. The actual speech of a character provides important characterisation cues, but equally the patterns that accompany the presentation of speech are of relevance. Ruano San Segundo (2016) explores reporting verbs in a corpus of Dickens novels and shows that some of the reporting verbs are used consistently with particular characters thus contributing to the individualisation of characters. Reporting verbs can also have stylistic functions in the narrative, for instance as "reminders" of characters who figure less prominently than the main characters for whom the author has a wider range of possibilities for characterisation. Ruano San Segundo's (2016) focus is on speech verbs other than *said*, which he considers "neutral" and not playing "a characterising role by itself" (Ruano San Segundo 2016: 116). However, as Ruano San Segundo also points out, *said* often comes "glossed" with further information, such as the manner of speaking (Caldas-Coulthard 1988: 167) so that it provides "the accuracy and range of details supplied by more specific choices of verbs" (Ruano San Segundo 2016: 117).

The verb form *said* is generally the most frequent reporting verb in English fiction and as such a more general reporting verb than *shouted* or *grumbled*, for instance. The repeated occurrence of *said* can invite the translator to vary its translation equivalents throughout a text. Reasons for such variation may lie in the literary conventions of the target language, as Levý (2011: 113) points out:

> Most professional translators are aware these days that the stereotypical repetition of *said* in English introducing direct speech quite simply belongs to a different literary convention, and as a rule they vary the way they represent this reporting verb in translation.

Employing a set of different options for the translation of *said* can potentially lead to subtle shifts in meaning and hence affect the presentation of a fictional character. Corness (2009) investigated translations of *said* in a parallel corpus of 22 English source novels translated into Czech. Examining nearly 10,000 occurrences of the reporting verb *said*, he found that it was translated by no less than 1,323 different Czech translation equivalents. Only 136 of the translation equivalents occurred 10 times or more, which indicates high creativity in the translations. He further reported that only in 33.3% cases a similarly neutral reporting verb was used in the Czech translation. Fárová (2016) studied reporting verbs in a parallel corpus of fiction containing translations between English, Czech and Finnish. In her English subcorpus (around 8.5 million words) she identified 168 different reporting verbs (i.e. word types). Of all the 5,053 occurrences of the reporting verbs (word tokens), *said* makes up 84%. Fárová (2016) then used another smaller translation corpus (0.4 million words, 6 English novels) and examined the translations of reporting verbs. The Finnish translators kept the nearest equivalent (*sanoi*) in the majority of cases (between 81% to 99%), but there was a clearer tendency in Czech for variation; the translators kept the nearest equivalent *řekl/řekla* on average in 70% of the cases (ranging between 40% to 91%). Fárová (2016) also mentions a clear departure from the norm: one of the translations showed extreme values, the closest equivalent *řekl/řekla* was kept in only 8% of the occurrences of *said* (out of 191) and the translator used 164 different equivalents. The tendency of Czech translators to avoid repetition and especially the repetition of the reporting verb *said* is also confirmed by Nádvorníková (2017) who examined translations of reporting verbs between Czech, French and English using parallel aligned corpora (i.e. the source text is aligned to its translation, in this case at sentence level). Fárová (2016: 148–149) further notes that the occurrence of *řekl/řekla* is somewhat higher in non-translated Czech texts. She examined a comparable corpus of translated vs. non-translated Czech fiction. While *řekl/řekla* ('said') is the most frequent reporting verb in both corpora, in non-translated Czech fiction it makes up 34% of all reporting verbs Fárová (2016) identified, while in the translated texts it only covers 25.5% of the reporting verbs.

3. *Said* and other reporting verbs in *Alice's Adventures in Wonderland*

The only previous corpus-based study of the *Alice* texts we are aware of is by Inaki and Okita (2006). In this study, the authors show the development of *Alice*'s character from a more passive *Alice* in *Alice's Adventures in Wonderland* to a more active heroine in *Through the Looking-Glass*. One of the textual features that Inaki and Okita (2006) draw attention to is the frequent use of reporting verbs, especially the verb *said*. They observe that while "Alice in *Wonderland* plays her role passively, which means *Alice* is in the position of being questioned and then responds to questions asked" (Inaki & Okita 2006: 286), her role during the story (and then subsequently in the second book) "changes radically from the responding type […] to the questioning type" (Inaki & Okita 2006: 287). Along with reporting verbs, Inaki and Okita (2006) also study adverbials modifying verbs of saying. In this section, we first look at overall frequencies of the reporting verb *said* in a larger corpus of contemporary fiction (contemporary to the *Alice* books). We then explore the text of *Alice's Adventures in Wonderland* and its Czech translation by Aloys Skoumal and Hana Skoumalová (published in 1961). We have chosen this highly acclaimed translation because it seems to be the most widely read one. We will focus on speech verbs reporting the speech of *Alice* and *the Queen*. In line with existing research on frequent translation shifts in reporting verbs in Czech translations (as discussed in Section 2), we will investigate these shifts in view of the role they play in depicting the characters *Alice* and *the Queen*. Section 4.1 will focus on *Alice* and Section 4.2 on *the Queen*.

With 462 occurrences, *said* is the most frequent reporting verb in the text of *Alice's Adventures in Wonderland*, followed by the much less frequent *went on* and *thought*[2] (48 and 43 occurrences respectively). Inaki and Okita (2006) observe that *said* is not only the most frequent reporting verb in this text, but its frequency is unusually high. This is confirmed when we compare the normalised frequencies of *said* in the *Alice* text with a selection of other 19th century books,[3] see Table 1.

The relative frequency of *said* in *Alice's Adventures in Wonderland*, 17,314 occurrences per million words, is relatively higher compared to most other books in Table 1. For the 16 fictional works we looked at (five of which are children's books), the normalised frequency of *said* ranges from 3,019 to 10,225. However, there is also one other work, where *said* is even more frequent than in *Alice* – Nesbit's *The Railway Children* with a frequency of 19,192. Both *Alice* and *The Railway Children* are full of conversation, so this finding is not unexpected.

2. Occurring as a reporting verb.

3. We use the online tool CLiC <http://clic.bham.ac.uk> (15 January 2018).

Table 1. Relative frequencies of the reporting verb *said* in *Alice* and other 19th century fiction (normalised per million words)

	Frequency of *said* per million words
Alice's Adventures in Wonderland	**17,445**
Wuthering Heights	3,218
Vivian Grey	3,553
Vanity Fair	4,430
The Woman in White	3,527
The Strange Case of Dr Jekyll and Mr Hide	5,079
The Small House at Allington	8,025
The Return of the Native	6,773
The Picture of Dorian Gray	3,329
The Mill on the Floss	7,956
Tess of the D'Urbervilles	3,996
Emma	3,019
Peter Pan	7,567
The Coral Island	4,201
The Jungle Book	8,415
The Railway Children	**19,192**
The Secret Garden	10,225

As corpus linguistics generally shows, frequency of occurrence is typically linked to functional patterns. An analysis of the concordance lines of the reporting verb *said* in *Alice* reveals three main patterns:

(a) *said* is accompanied by a description of following or simultaneous action (Examples (1) and (2)). When *said* (but also other verbs indicating speech) is followed by such a description of action, the speech functions as a point of transition in the narrative. This kind of narrative progression has also been observed by Toolan (2009) and Busse (2010).

(1) she said to herself as she ran

(2) As she said this, she came upon a neat little house

(b) *said* is followed by body language (see Examples (3) and (4));

(3) the Hatter said, tossing his head contemptuously

(4) the Cat said, waving its right paw round

(c) *said* is accompanied by additional information describing the manner of speaking (see Examples (5) and (6)) or describing the character who is speaking (Example 7).

(5) said the Caterpillar sternly

(6) she said to herself in a melancholy tone

(7) said Alice, who was a very truthful child

Especially patterns of body language (b) and descriptions of the manner of speaking (c) show the presence of the narrator who imposes control over the text "through the insertion of reporting clauses accompanying the stretches of [direct speech]" that inform the reader of an "additional note of attitude" (Busse 2010). Textual patterns of body language and descriptions of manner of speaking are crucial features for the modelling of fictional characters (cf. Stockwell & Mahlberg 2015: 134). Mahlberg and Smith (2010) and Mahlberg (2013) have found patterns similar to those illustrated here by focusing on 'suspensions', i.e. stretches of narrator text that interrupt the speech of a character, mostly in the form of a reporting clause, as in Example (8), where the suspension is italicised.

(8) 'It IS a long tail, certainly,' *said Alice, looking down with wonder at the Mouse's tail*; 'but why do you call it sad?'

Because of the close link of direct speech, body language and the manner of speaking, suspensions are a particularly useful places to look for subtle character hints. In the present analysis, however, we cast the net wider and include all types of clauses with *said*. So (5), given below as (5a) with more context, is not a suspension as the clause ends with a full stop and does not continue with quoted speech. The sentence in Example (2) above does not even include direct speech.

(5a) 'What do you mean by that?' said the Caterpillar sternly.

For the 462 occurrences of *said*, the most frequent subject is *Alice* (36%), followed by *the King* (9%), *the Mock Turtle* (7.5%), *the Hatter* (4.8%), *the Queen* (4.5%), and *the Gryphon* (4%). If we compare these figures with the raw frequencies of the character names, we notice, however, that while *Alice* is the most frequent character name in the novel (with 397 occurrences), the second most frequent name is not *the King* but *the Queen* (with 72 occurrences), see Table 2.

Table 2. The most frequently occurring characters in *Alice's Adventures in Wonderland*. Column 2 indicates the number of occurrences of the character name in the text, column 3 shows the number of instances where the character name occurs in the same sentence as the reporting verb. In the majority of the cases the reporting verbs occur in the simple past tense but we also count other forms. Column 4 shows how many times the character speech was introduces by *said*.

Character	Occurrences	Speech	with *said*
Alice	397	218	130 (59.6%)
the Queen	72	32	18 (56.3%)
the King	62	47	39 (83%)
the Mock Turtle	52	37	24 (64.9%)
the Gryphon	52	44	19 (43.2%)
the Hatter	52	37	22 (59.5%)

Table 2 shows that *said* is the most frequent reporting verb for all the characters listed; for *the King said* introduces most of the instances of him speaking (83%). Similar to Ruano San Segundo's (2016) findings on reporting verbs, some of the more expressive and less frequent reporting verbs are typically associated with specific characters, e.g. *screamed* is used with *the Queen, the Gryphon* and *the Pigeon, pleaded* and *thought* is used exclusively with *Alice*. Table 3 shows the distribution of the reporting verbs for the six most frequently mentioned characters (as shown in Table 2).

The main verbs for *Alice* are *said* (59.6%), *thought, replied*, and *began*. The main reporting verb for *the Queen* is also *said* with a similar proportion, i.e. 56.3%, her second most frequent verb is *shouted* (the count includes the form *shouting*) and most of the additional reporting verbs are more expressive, indicating loudness and anger (e.g. *screamed, roared*). Similarly, for *the King* the main reporting verb is *said* (83%) and the rest of the verbs (all occurring just once) are rather neutral (e.g. *replied, added, repeated*), except for *exclaimed* and *persisted*. The situation is similar for *the Mock Turtle* and *the Hatter*, for whom *said* is the most frequent verb and more expressive verbs occur just once (e.g. *sighed, muttered, grumbled*). *The Gryphon* seems to be a more vivid character: there is a whole range of verbs, from the most frequent *said*, but with lower frequency compared to other characters (43.2%) and other neutral reporting verbs (e.g. *went on, added*) to verbs such as *shouted, screamed, yelled*, and *whispered* (all occurring once). Based on these results, *the Queen*'s character has a clearly individualised profile in terms of how her speech is introduced. In addition to reporting verbs, we also find additional descriptive information on her manner of speech, as in Examples (9) and (10).

Table 3. Reporting verbs of the six most frequent characters

Alice		***the Queen***		***the King***		***the Mock Turtle***		***the Hatter***		***the Gryphon***	
***said*(59.6%)**	**130**	***said*(56.3%)**	**18**	***said*(83%)**	**40**	***said*(64.9%)**	**24**	***said*(59.5%)**	**22**	***said*(43.2%)**	**19**
*thought**	33	***shouted***	7	*replied*	1	*went (on)*	3	*went (on)*	4	*went (on)*	3
replied	14	*added*	2	*added*	1	*replied*	3	*added*	2	*added*	3
began	6	*screamed*	2	*called (out)*	1	*cried (out)*	2	*continued*	2	*cried (out)*	3
cried (out)	6	*shrieked out*	1	*remarked*	1	*began*	1	*replied*	1	*replied*	2
asked	5	*roared*	1	*repeated*	1	*asked*	1	*began*	1	*interrupted*	2
went (on)	4	*bawled out*	1	*continued*	1	*repeated*	1	*asked*	1	*answered*	2
remarked	4			*exclaimed*	1	*interrupted*	1	*interrupted*	1	no reporting verb	2
exclaimed	3			*persisted*	1	*persisted*	1	*sighed*	1	*asked*	1
pleaded	3							*muttered*	1	*remarked*	1
called (out)	2							*grumbled*	1	*repeated*	1
whispered	2									*shouted*	1
added	1									*continued*	1
interrupted	1									*whispered*	1
shouted	1									*screamed*	1
answered	1									*yelled*	1
sighed	1										
inquired	1										
panted	1										

The counts are based on the number of concordance lines for character names where the character name occurs together with the reporting verb in the same sentence. The most frequent verb form is listed in the table, but other forms are also included in the counts, e.g. *Alice ventured to remark* is counted under *remarked*.

(9) as she heard the Queen's voice in the distance, screaming with passion

(10) and the shrill voice of the Queen ordering off her unfortunate guests to execution

In contrast to *the Queen*, *Alice* has a wide variety of verbs that report her speech. What sets her apart from the other characters in the text is the reporting verb *thought* and specifically *thought to herself* indicating the importance of her inner speech (see Section 4.1.1). The verb THINK (as a lemma) occurs 133 times in the novel and it is overwhelmingly *Alice* who does the thinking; she even asserts explicitly: *I've a right to think*. Of the 133 occurrences of THINK, there are only 19 that do not refer to *Alice*. These include examples of *the Dormouse* or Alice's sister, or uses of the verb in *as if* constructions (*as if it thought*). The emphasis on Alice's ability and right to think seems to be in line with readings of the *Alice* books as "portrayals of the experience of growing up and the construction of agency and identity" (Sigler 1997: xiv). At the same time, Alice's inner thoughts are revealed as she is the main focaliser. In Section 4, we will examine the individual translation solutions for *Alice* and *the Queen* – the two most frequently mentioned characters, focusing particularly on the characterising nature of the way Carroll introduces their speech and how this is rendered by the translator's voice.

This overview of characters and reporting verbs shows how the frequency with which specific reporting verbs are used can be tied to individual characters to the extent that some reporting patterns are exclusive to individual characters. Additionally, the patterns of *said* illustrate that character information will not automatically be signalled by the reporting verb alone, but the verb is part of larger patterns, as seen in Examples (1) to (8).

4. Translating *Alice* and *the Queen*

Table 4 shows reporting verbs that occur with *Alenka* [Alice] and *královna* [the Queen] in the Czech translation. Only those occurrences where the name is explicitly mentioned have been counted[4] and the table only lists verbs that occur at least twice. For *Alenka* there are additionally 35 verbs that occur only once (*div nevybuchla*, *doprošovala se*, *dorážela*, *hartusila*, *hlesla*, *libovala si*, *myslila si*, *nadurdila se*, *oddechla*, *odpovídala si*, *podotkla*, *polekala se*, *pošeptala*, *pravila*, *prosila*, *říkala si*, *rozkřikla se*, *rozplakala*, *škemrala*, *svitlo*, *troufla si*, *varovala*, *volala*, *vpadla*

4. We have used the *InterCorp* parallel corpus, which includes both the English original of *Alice's Adventures in Wonderland* and the translation, available at <www.korpus.cz> (15 January 2018).

mu do řeči, *vybuchla*, *vyhrkla*, *vykládala*, *vypravila ze sebe*, *žadonila*, *zahořekovala*, *zamlouvala*, *zamyslela se*, *zašeptala*, *zavolala, žasla*). For *královna* there are 15 verbs that occur just once (*kázala, křičela, okřikla, podotkla, pokřikovala, přerušila, prohodila, řekla, rozkřikla se, rozlítila se, spustila, zaječela, zařvala, zavřískala, zeptala se*).

Table 4. Reporting verbs used with *Alenka* [Alice] and *královna* [the Queen] in the Czech translation (based on the concordance of *Alenka* and *královna* respectively, the table only includes verbs occurring at least twice)

***Alenka* [Alice]**		***královna* [the Queen]**	
řekla	72	*křikla*	3
řekla si	22	no verb	2
pomyslila si	15		
zeptala se	13		
odpověděla	10		
bránila se	5		
no verb	5		
vyhrkla	5		
spustila	4		
odsekla	3		
přemýšlela	3		
pronesla	3		
zvolala	3		
chlácholila	2		
křikla	2		
napadlo	2		
oddechla si	2		
okřikla	2		
osmělila se	2		
posteskla si	2		
vykřikla	2		
zlobila se	2		

What is immediately obvious from Table 4 and the range of reporting verbs that occur just once is the great variety of the verbs, confirming earlier findings as discussed in Section 2. While in the English original, there are 19 different reporting verbs introducing *Alice*'s speech (see Table 3), there are 56 different reporting

verbs used in the Czech translation. Similarly, there are only seven different reporting verbs used to introduce or comment on *the Queen's* speech, while there are 16 different verbs in Czech. Even more remarkably, the main reporting verb *řekla*, corresponding to the English *said*, is used to report *Alice*'s speech in 43.5% of the cases (this includes the reflexive variant *řekla si* [said to herself]) but with the speech of *the Queen* it occurs just once.

4.1 *Alice*

We analyse *Alice* in the English original in Section 4.1.1 and Section 4.1.2 will deal with the Czech translation.

4.1.1 Alice *in English*

Alice is a well-mannered, educated, "confused Victorian schoolgirl" (Horton 2002: 99). Based on their corpus analysis, Inaki and Okita (2006: 291) conclude that *Alice* is "almost always on the defensive and annoyed". Linguistically, however, the picture seems to be more complex, as a selection of concordance lines for the character *Alice* shows:

'I don't know of any that do,'	**Alice**	said very politely, feeling quite
think I should understand that better,'	**Alice**	said very politely, "if I had it
'Please would you tell me,' said	**Alice**,	a little timidly, for she was not
'A knot!' said	**Alice**,	always ready to make herself useful
'Is that all?' said	**Alice**,	swallowing down her anger as well
'I – I'm a little girl', said	**Alice**,	rather doubtfully, as she
very civil of you to offer it', said	**Alice**	angrily. 'It wasn't very civil of
beautify is, I suppose?' 'Yes,' said	**Alice**	doubtfully: 'it means to make
'As wet as ever,' said	**Alice**	in a melancholy tone:
'Well, perhaps not,' said	**Alice**	in a soothing tone:
'Only a thimble,' said	**Alice**	sadly. 'Hand it over here,' said

Based on the concordance of *Alice* (397 occurrences), it is possible to identify the reporting verbs that are used with this character. Table 3 in Section 3 shows an overview of these reporting verbs for the six most frequently mentioned characters. The verb *said* is the most frequent one (occurring 130 times with *Alice*[5]), as it is for all the characters. In about a third of the cases the verb *said* occurs further

5. The figure 130, cf. Table 3, includes all instances of the verb SAY, i.e. not only the form *said*, which occurs 127 times but also instances like: *Alice was very nearly getting up and* ***saying****….*

modified by the manner of speaking most frequently in the form of an *–ly*-adverb of manner:

touch her. 'Poor little thing!' **said Alice**, **in a coaxing tone**, and she tried
'I – I' m a little girl,' **said Alice**, rather **doubtfully**, as she
with her head! Off – ' 'Nonsense!' **said Alice**, very **loudly** and **decidedly**, and
very civil of you to offer it,' **said Alice** **angrily**. 'It wasn't very civil of
beautify is, I suppose?' 'Yes,' **said Alice** **doubtfully**: 'it means to make
'Come, my head's free at last!' **said Alice** **in a tone of delight**, which change
'There's plenty of room!' **said Alice** **indignantly**,and she sat down in a
'I've a right to think,' **said Alice** **sharply**, for she was beginning to
'That would be grand, certainly,' **said Alice** **thoughtfully**: 'but then – I should
'I beg your pardon,' **said Alice** very **humbly**: 'you had got to the

Other modifications include a description of *Alice*'s effort to change the topic of the conversation (*in a hurry to change the subject*), descriptions of *Alice*'s state of mind (*a good deal frightened, feeling very glad, rather alarmed by the proposal*), descriptions of simultaneous or following action (*as she swam about, as she picked her way through the wood*), and there is also one occurrence of accompanying body language description (*looking down with wonder at the Mouse's tail*, see Example 8 in Section 3):

name again!' 'I won't indeed!' **said Alice**, in a great hurry to change the subject
'Really, now you ask me,' **said Alice**, very much confused, "I don't think – "
'The was a narrow escape!' **said Alice**, a good deal frightened at the sudden
'I wish I hadn't cried so much!' **said Alice**, as she swam about, trying to find her

Out of the 127 occurrences of *said*, in four cases *Alice* is speaking to herself (*said to herself*). The second most frequent reporting verb is *thought* (occurring altogether 33 times, cf. Table 3),[6] which introduces inner speech and occurs exclusively with *Alice*. Out of the 33 occurrences, *thought* occurs five times together with *to herself*, as in Examples (11) and (12).

(11) 'How fond she is of finding morals in things!' **Alice thought to herself**.

(12) 'One side of what? The other side of what?' **thought Alice to herself**. 'Of the mushroom,' said the Caterpillar

6. Occurrences of cognitive meanings of this verb were excluded, e.g. *Alice thought the whole thing very absurd*.

The phrase *to herself* occurs exclusively with *Alice* (46 times) as the focaliser of the narrative, stressing the importance of her inner state of mind (see Section 3). In addition to the five occurrences with *thought/ think* and 30 occurrences with *said/ saying* the phrase occurs with seven other verbs (*planning, went on, talking, muttered, fancy, added, pictured*). Speaking and thinking to herself is a crucial feature of *Alice*. It makes her stand out against the other characters. The fact that this distinction between *Alice* and the inhabitants of the Wonderland is made through the use of reporting verbs is in line with the nonsense conversation that characterises the Wonderland. *Alice* comes from a different world, her manner of speaking and thinking differs from the rules of conversation in the Wonderland.

4.1.2 Alice *in Czech*

In the Czech translation *Alice* seems to be much more emotional than in the original text. Her emotions range from scared (13), confused (14), to annoyed (15), brave (16), and even enjoying herself (17).

(13) a Alenka samým strachem ani necekla
[and Alice did not dare to speak out of fear]

(14) Alenka nevěděla, co počít a v té bezradnosti sáhla do kapsy
[Alice did not know what to do, and in despair she put her hand in her pocket]

(15) to nebylo Alence po chuti
[and Alice did not like that]

(16) sama užasla, že si tak troufá
[she was astonished herself that she was so daring]

(17) nenuceně bavila
[she chatted spontaneously]

The range of emotions is also clearly visible in the premodifying adjectives and nouns that accompany the name. While in the English original we only have *poor* (10 times), *little* (twice), and *wise little*,[7] there is a greater range of premodifiers (adjectives and a noun) in Czech: *chudák* (twice) [poor], *moudrá* [wise] (once),

7. The collocation *wise little* is not generally frequent (there are no occurrences in the BNC) but we can find it in children's discourse such as *Wise little Hen* by Walt Disney. In the 4.4 million-word *19th Century Children's Literature Corpus* ChiLit (which was compiled as part of the GLARE project <https://www.birmingham.ac.uk/schools/edacs/departments/englishlanguage/research/projects/glare/index.aspx> supported by the Marie Curie Research Grants Scheme ref. EU 749521; and is available at <http://clic.bham.ac.uk/> (15 January 2018)) it occurs only three times (once in *Alice's Adventures in Wonderland* as discussed above).

nešťastná (5 times) [unhappy], *celá šťastná* [so happy] (once). This range is partly due to the fact that Czech allows a wider range of premodification of proper names than English does. Overall, *Alice* in Czech is more vivid and exaggerated (Examples (18) to (21)) than her smooth, polite, Victorian original.

(18) 'Don't grunt,' said Alice; 'that's not at all a proper way of expressing yourself.'
„Nechrochtej," okřikla je Alenka, „tak se přece nemluví."
['Don't grunt,' Alice snapped at them, 'this is not a proper way to speak.']

(19) 'You should learn not to make personal remarks,' Alice said with some severity;
„Co si to dovolujete," okřikla ho Alenka.
['How do you dare,' Alice snapped at him.]

(20) 'Don't talk nonsense,' said Alice more boldly: 'you know you're growing too.'
„Nemluvte hlouposti," spustila zostra Alenka, „však vy rostete taky."
['Don't talk nonsense,' Alice began sharply, 'you're growing too after all'.]

(21) said Alice, swallowing down her anger as well as she could.
Alenka div nevybuchla.
[Alice nearly exploded.]

The above observations are based on a concordance of *Alice*. The focus on the translations of the reporting verb *said* makes the picture more precise. Out of the 115 occurrences of *said Alice*, only 67 are translated with the verb *řekla*, which is the nearest equivalent. This means 42% (48 occurrences) of *said* are translated in some other way, see Table 5 for translation solutions.

Table 5. Translation solutions for the phrase *said Alice* (115 occurrences)

Translation solution	
said	67 (58%)
Specific reporting verb	22 (19%)
reply/ask	11 (10%)
Omission	6 (5%)
Translator creativity	5 (4%)
Translation shift	4 (3%)

Note: *said* = *řekla*, Specific reporting verb = reporting verb is specific but is in line with the original text; *odpovědět* = *reply*/*zeptat* = *ask*; Omission = the reporting verb (*said*) is omitted in the translation; Translator creativity = the translator chooses a creative solution with no substantial shift in meaning; Translation shift = there is a shift in meaning.

Most of the cases where the translators have chosen a reporting verb that is different from *said* are in line with the original text and conform to the Czech stylistic tradition of avoiding repetition, see Examples (22) to (24).

(22) 'I wish you wouldn't squeeze so,' said the Dormouse, who was sitting next to her. 'I can hardly breathe.'
'I can't help it,' **said** Alice **very meekly**: 'I'm growing.'
Plch seděl vedle ní a hned se ozval: „Netlačte se tak na mě. Sotva dýchám"
„Já za to nemohu," **chlácholila** ho Alenka. „Já rostu."
[The Dormouse was sitting next to her and immediately responded: 'Don't push on me so hard. I can barely breathe.'
'I can't help it,' Alice **was soothing** him. 'I'm growing.']

(23) 'Thank you, it's a very interesting dance to watch,' **said** Alice, **feeling very glad** that it was over at last:
„Děkuji, napohled je ten tanec velmi zajímavý," Alenka **si oddechla**, že už je konečně po něm,
['Thank you, the dance is very interesting to watch,' Alice **was relieved**, that it was finally over,]

(24) 'Wouldn't it really?' **said** Alice **in a tone of great surprise**.
„Vážně?" **žasla** Alenka.
['Really?' Alice **was astonished**.]

In Example (22), *said Alice very meekly* is translated as *chlácholila ho Alenka* [Alice was soothing him]. The Czech verb *chlácholit* is not a typical reporting verb and means "soothe" both verbally and physically and is often used with babies, small children, or somebody who is very upset or crying. In Examples (23) and (24), the Czech translation again omits direct reporting verbs and instead describes the emotional state of *Alice* ("relieved", "astonished"). These verbs are, however, occasionally used in place of reporting verbs in Czech.

In 12.5% of the cases, the translators choose not to translate the reporting *said* at all (which is not unusual in Czech), see Examples (25) and (26).

(25) 'A knot!' said Alice, always ready to make herself useful, and looking anxiously about her.
„Smyčku!" Alenka ráda každému pomáhala, a hned se tedy ohlížela kolem sebe.
['A knot!' Alice was happy to help everyone, so she was looking around herself immediately.]

(26) 'If you're going to turn into a pig, my dear,' said Alice, seriously, 'I'll have nothing more to do with you. Mind now!'
„Jestli se, milánku, proměníš v prasátko, už s tebou nechci nic mít. Dej si pozor!"
['If you are, going to turn into a pig, my dear, I'll have nothing more to do with you. Mind you!']

There are several instances where we come across greater translator creativity conveying the translators' interpretation, as in Example (27), where the simple *said Alice* is rendered by the expressive *bránila se Alenka* [Alice was defending herself] and Example (18 above), where the simple *said Alice* is translated as *okřikla je Alenka* [Alice snapped at them].

(27) 'If you knew Time as well as I do,' said the Hatter, 'you wouldn't talk about wasting it. It's him.'
'I don't know what you mean,' **said** Alice.
„Kdybys znala Čas tak jako já," pravil Kloboučník, „nemluvila bys o maření času. Je to někdo."
„Já vás nechápu," **bránila se** Alenka.
['If you knew Time as well as I do,' said the Hatter, 'you wouldn't talk about wasting time. It is someone.'
'I don't understand you,' Alice **was defending herself**.]

As discussed in Section 4.1.1, the characterising reporting verb for *Alice* is *thought*. With *Alice, thought* occurs 42 times (this includes occurrences where she is referred to as *she*, hence a different number than discussed in Section 4.1.1, where only occurrences with *Alice* were considered). The translators choose a range of translation solutions: the most frequent equivalents they use are *řekla si* [said to herself] and *pomyslila si* [thought to herself], both occurring 15 times. Other translation solutions include *napadlo* [occur], *přemýšlela* [was thinking], *oddychla si* [was relieved] and *řekla* [said]. Only *oddychla si* and *řekla* do not report inner speech and thought. The two main equivalents the translators use for reporting inner speech and thought *řekla si* and *pomyslila si* are used exclusively with Alice.

4.2 The Queen

In this Section, 4.2.1 looks at the source text and 4.2.2 deals with the target text.

4.2.1 The Queen *in English*

The phrase *the Queen* occurs altogether 72 times, in 32 of these cases the Queen speaks (see Table 3 above). A characteristic feature of *the Queen* is the cluster *Off with [his/her/their] head* that she utters altogether ten times (Example 28).

(28) The Queen had only one way of settling all difficulties, great or small. 'Off with his head' she said, without even looking round.

The only adjective modifying the *Queen* is *savage*. *The Queen* presents herself in a loud and angry manner (*shouting*, *screaming*, *roared*, *voice of thunder*) and with expressive body language (*stamping on the ground, turned crimson with fury, frowning like a thunderstorm*). All this makes her an exaggeratedly terrifying figure.

The choice of reporting verbs provides a similar picture. Even though *said* is the most frequent reporting verb (18 occurrences), there are also more expressive verbs that occur with *the Queen: shouted* (five times and *shouting* twice), *screamed* (twice) *bawled out, shrieked out*, and *roared* (one occurrence each):

'Come on, then!' **roared the Queen**, and Alice joined the procession, wondering
'Leave off that!' **screamed the Queen**. 'You make me giddy.' And then, turning to
'That's right!' **shouted the Queen**. 'Can you play croquet?' The soldiers were
Get to your places!' **shouted the Queen** in a voice of thunder, and people began
you fair warning,' **shouted the Queen**, stamping on the ground as she spoke;
their heads off?' **shouted the Queen**. "Their heads are gone, if it please your
said the Hatter, when **the Queen** jumped up and **bawled out**, 'He's murdering
playing **the Queen** never left off quarrelling with the other players, and **shouting**
'Collar that Dormouse!' **the Queen shrieked out**. 'Behead that Dormouse! Turn
'Off with her head!' **the Queen shouted** at the top of her voice. Nobody

The phrase *said the Queen* occurs 12 times (while the phrase in the reversed word order *the Queen said* occurs four times) and in half of the cases this phrase is further modified, in two cases the manner of speaking is specified (see Examples 29 and 30):

(29) 'Get up!' said the Queen, in a shrill, loud voice, and the three gardeners instantly jumped up,

(30) 'Never!' said the Queen furiously, throwing an inkstand at the Lizard as she spoke.

In two cases, the body language that accompanies her speech is described (Examples 31 and 32):

(31) 'Idiot!' said the Queen, tossing her head impatiently; and, turning to Alice, she went on, 'What's your name, child?'

(32) 'Hold your tongue!' said the Queen, turning purple. 'I won't!' said Alice. 'Off with her head!' the Queen shouted at the top of her voice.

In the remaining two cases simultaneous action is described (*pointing to the three gardeners who were lying round the rose-tree; who had meanwhile been examining the roses*).

Even though the reporting verb *shouted* also marginally occurs with other characters, it is the most frequent verb (except *said*) for *the Queen*. Considered together with the other "loud" verbs that we find for her, it is one of her individualising verbs. When we examine the verb *shouted* in a bigger reference corpus of 19th century children's literature (the ChiLit corpus, which contains 4.4 million words across 71 books, see also Note 7), we notice that this verb is typically used with male characters, while it is only used with female characters in about 4% of its occurrences. A similar observation can be made for the verb *roared*, that occurs with *the Queen*, too. The verb *roared* is used with female characters in about 3% of its occurrences, see Table 6. Considering the various literary interpretations of the Wonderland's characters and gender dynamics in the book, we should note that *the King*'s speech is, in contrast to that of his *Queen*, reported mainly by the verb *said* and other neutral reporting verbs as discussed in the previous section (cf. Table 3).

Table 6. The verbs *shouted* and *roared* in the ChiLit corpus and the gender of the character whose speech is reported by these verbs.

	shouted	***roared***
Male	358 (78%)	59 (60%)
Female	20 (4%)	3 (3%)
Not enough context	83 (18%)	5 (5%)
Inanimate subject		32 (32%)

Note: 'Not enough context' indicates that the concordance lines did not give enough context to determine the speaker's gender.

4.2.2 The Queen *in Czech*

The image of *the Queen* (*královna*) as a loud, angry and terrifying character is portrayed very similarly in the Czech target text. However, there seems to be a slight shift, since there is a sense of dignity about *the Queen* in the English original, while in Czech she sounds impulsive and much less serious (Example 33):

(33) said the Queen, in a shrill, loud voice
rozkřikla se na ně zhurta Královna
[the Queen started screaming at them impulsively]

There are other differences worth further consideration. As pointed out in Section 3, the use of *said* in Carroll's text clearly supports the narrator's voice in different ways. This is most apparent in body language descriptions and

descriptions that add detail to the manner of speaking. Such descriptions reflect the speaker's attitude and at the same time make the character vivid for the reader by highlighting for example the acoustic quality of the voice, see Example 34. It is therefore arguable what we gain or lose in translation solutions where a clear shift in meaning of the reporting verb is present.

(34) the Queen said severely
Královna se rozkřikla
[the Queen started screaming]

When we examine the translation solutions of the phrase *said the Queen* in a similar way as we did for *Alice*, it turns out that in half of the cases *said* is not translated by its nearest equivalent *řekla*, see for instance Examples (35) and (36).

(35) 'I see!' **said** the Queen, who had meanwhile been examining the roses. 'Off with their heads!' and the procession moved on
Královna si už zatím prohlédla růže a **přerušila** ho: „No bodejť! Hlavu jim srazit!" Průvod se hnul kupředu
[The Queen has meanwhile examined the roses and **interrupted** him: 'Well then! Off with their heads!' The procession moved on]

(36) 'Hold your tongue!' **said** the Queen, turning purple.
„Budeš mlčet!" **okřikla** ji Královna a celá zbrunátněla.
['Will you shut up!' the Queen **snapped** at her and turned all purple.]

There is an apparent translation strategy to avoid the repetitive *said*. Instead, the translators choose a great variety of expressive verbs. Most often the equivalents are based on the verb *křičet* [shout, scream] and its variants, where the meaning is modified by the prefix and/or verb aspect: *rozkřiknout se, křiknout, pokřikovat, okřiknout*. There are also several other more or less synonymous verbs *(zavříškat, zaječet, zařvat*). So, despite avoiding the repetition of *řekla* [said], the Czech translation arguably still achieves a repetitive impression by using a range of verbs that create similar meanings with the majority of them being based on the same word stem *křik-*.

Additionally, in the English text, we found the uses of *shouted* and *roared* to be gendered, which is less so the case for the Czech equivalents based on the verb *křičet* frequently used with *the Queen*. There are two main equivalents of *shout* in Czech, these are *křičet* and *řvát*, the latter is stronger and closer to the English *roar*. An analysis based on a corpus of children's literature written by Czech authors[8]

8. This corpus was created on 26 September 2016 as a subcorpus from the large SYN corpus of contemporary Czech <http://wiki.korpus.cz/doku.php/en:cnk:syn>. The subcorpus includes fiction texts where the original language is Czech (i.e. translations are not included) and which

shows that both of these verbs (and their variants) tend to be used more frequently with male characters, but both are still used with female characters (of the two, *křičet* is used with female characters more often than *řvát*). The choice of these equivalents further indicates the translators' tendency for normalisation. Since there are so many variants of these two base verbs in Czech, this point requires more detailed analysis. Therefore, we have analysed two verbs – *zařvat* (a variant of *řvát*) and *křiknout* (a variant of *křičet*) – in the reference corpus of Czech children's literature, see Table 7. The verb *zařvat* occurs once in the translation and it is *the Queen* whose speech is reported with this verb. The source text contains the verb *roar*, see Example (37).

Table 7. *Křiknout* ['shout' – perfective aspect indicating short duration] and *zařvat* ['shout/roar' – perfective aspect indicating short duration] and the gender of the character whose speech is reported

	křiknout	***zařvat***
Male	57 (69%)	109 (69%)
Female	23 (**28%**)	21 (**13%**)
Other	3 (4%)	27 (17%)

(37) 'Yes!' **shouted Alice**. 'Come on, then!' **roared the Queen**, and Alice joined the procession, wondering very much what would happen next.
„Já," **křikla Alenka**. „Tak pojď!" **zařvala na ni Královna**; Alenka se přidala k průvodu a trnula, co bude dál.
['Me,' **shouted** Alice. 'Come on, then!' **roared** the Queen at her; Alice joined the procession worried what would happen next.]

Example (37) depicts the encounter of *the Queen* and *Alice* before the croquet game with *Alice* standing up boldly to *the Queen*. This is reflected by the choice of reporting verbs. *Alice* replies to *the Queen* by 'shouting' and this is one of only two occurrences of *shouted* used with *Alice*. *The Queen* responds by 'roaring', a verb a degree louder than *Alice*'s shouting. This is the only example where *the Queen* roars – she shouts more frequently than she roars (see Table 3). The gradually upgraded loudness is important for the development of the plot, which is also reflected in the translation.

The main translation equivalents of *shout* are various variants of *křičet* (seven out of the eleven equivalents for *shouted*); the variant *křiknout* (in the perfective aspect) occurs four times and is thus the most frequent. However, the translators' use of the verb *křiknout* shows that this equivalent is used for various kinds of

are labelled as JUN (children's literature). The subcorpus contains 2.6 mil. words. All SYN corpora are available at <www.korpus.cz> (15 January 2018).

loud speech in the source text: *shout, cry (out), call out* and even *cheer*. The Czech verb therefore does not only indicate an angry and loud manner of speaking that intimidates the person who is being spoken to (Example 38 and 39) but is also used in various other "loud" situations (Examples 40 to 43).[9]

(38) in a very short time the Queen was in a furious passion, and went stamping about, and **shouting** 'Off with his head!'
netrvalo to dlouho, a Královna se tak rozlítila, že každou chvilku dupla a **křikla**: „Srazte mu hlavu!"
[it wasn't long and the Queen got so angry, that from moment to moment she stamped and shouted: 'Off with his head!']

(39) Now, I give you fair warning,' **shouted** the Queen, stamping on the ground as she spoke;
„Předem vás upozorňuji," Královna **křikla** a přitom dupla nohou,
['I warn you beforehand,' shouted the Queen and stamped her foot at the same time,]

(40) but she stopped herself hastily, for the White Rabbit **cried out**, 'Silence in the court!',
ale hned zas zmlkla; Králík totiž **křikl**: „Ticho v soudní síni!"
[but she immediately went quiet again; for the Rabbit shouted: 'Silence in the court!']

(41) and, when it had finished this short speech, they all **cheered**.
a když ten krátký proslov dopověděl, všichni **křikli** hurá.
[and when it finished the short speech, they all shouted hooray.]

(42) Here one of the guinea-pigs **cheered**, and was immediately suppressed by the officers of the court.
Jedno morče **křiklo**: „Výborně," a soudní zřízenci je rázem zlikvidovali.
[One guinea-pig shouted: 'Great,' and court officers immediately disposed of him.]

(43) The first thing she heard was a general chorus of 'There goes Bill!' then the Rabbit's voice alone-
Nejprve **křikli** sborem: „Vilík letí!," pak povelel Králík:
[First they shouted in a chorus: 'Bill is flying!,' and then the Rabbit commanded:]

9. The verb *křiknout* is used 13 times altogether. In four cases it corresponds to the source text *shout* (Examples 38 and 39), four times it corresponds to *cry (out)* (Example 40), three times it corresponds to *cheer* (Examples 41 and 42), once it corresponds to *call out* and once it does not correspond to a reporting verb at all (Example 43).

The translators' rather free handling of reporting verbs can also be seen in the following Example (44), where *řvát*, i.e. 'shout/roar', appears in Czech, while the source text contains the simple *said*.

(44) while the Mock Turtle sang this, very slowly and sadly: 'Will you walk a little faster?' **said** a whiting to a snail.
Paželv přitom zvolna a smutně prozpěvoval: „Hněte sebou, no tak honem," bělice **řve** na šneka.
[while the Mock Turtle sang slowly and sadly: 'Move on, come on fast,' a whiting roars at a snail.]

In the English original, *the Queen* is a terrifying figure. This impression is supported by the strong masculine reporting verbs used in the source text that make *the Queen* stand out from the rest of the characters. As we have shown above, this is only partially rendered in Czech. The reason is the lack of consistency in the translation so that the various translation equivalents create a rather diverse picture. Based on its frequency, *křiknout* is the most typical verb for *the Queen* but it is actually not a very "loud" verb, in addition, the perfective aspect stresses its short duration. Neither is it used with *the Queen* only. Its examination in a larger reference corpus (see Table 7) also shows that even though it is more commonly used with male characters, it still occurs with female characters. If the translators had chosen the verb *zařvat* instead, the gendered use of reporting verbs for *the Queen* could have been preserved more clearly. Based on the reference corpus check, the verb *zařvat* mainly appears with male characters, but in the *Alice* text, the translators only use it once.

For both *Alice* and *the Queen* reporting verbs provide crucial character information. While "loud" verbs, such as *screamed* and *shouted*, are typiclal of *the Queen*, with *Alice* we find inner speech (the phrase *to herself* occurs exclusively with *Alice*). The reporting verb *said* is the most frequent for all characters, but it still provides character information as part of reporting clause patterns. Because of its frequency, *said* offers a range of potential for translation and in line with the Czech stylistic tradition, the translators aim to avoid the repetitive use of *said*.

5. Conclusions

Our corpus linguistic framework for the translation of fictional characters is based on an innovative combination of process-, product- and function-oriented aspects of translation situated within the theoretical context of mind-modelling. Literary translation is a process – not only as translation, but also in the sense that it contructs a mental representation of a fictional character. The target text is

both the product of the translation and a literary reading of the source text. The function of the translation is to produce a text that can be read as a literary text and so its function also is to provide the reader with appropriate character cues. If we see *Alice's Adventures in Wonderland* as children's literature, we may also want to see the function of the translation as to produce a text aimed at children. The comparison of the source and target text that we provided in this chapter goes beyond the identification of features of translated texts or stylistic choices of the translators. The textual shifts that we identified are a reflection of the shifts in the modelling of the fictional characters. We have drawn on frequency comparisons of linguistic cues to compare different characters within the source and target text, but we have also compared translation equivalents across source and target texts and supported the assessment of the translations through comparisons with larger reference corpora, including a corpus of children's literature. In this way, we have not only described the evidential base of the fictional characters, our approach also takes the theory of mind-modelling beyond its monolingual version.

Differences between textual features in source and target text reflect that the translator's sense of person-ness is shaped linguistically, culturally and socially. Our analysis made this specifically apparent in the discussion of gendered uses of reporting verbs. As we have identified translation shifts with the help of corpus linguistic methods we cannot know to what extent the translators were aware of them. There are various potential reasons for the occurrence of translation shifts. As we have argued, the translation practice of avoiding repetition in Czech may be one. Another reason might be that a translator takes a narrow view, translating sentence by sentence rather than taking wider patterns across the text into account. We have shown that *said* in English is a reporting verb with rather general meaning, but it is still part of patterns that provide specific character information in the reporting clause. The range of verbs we have found in Czech to translate *said* introduce a shift in the verb meaning, but may add further shifts that affect the patterns around *said*. Finally, translation shifts may be the result of the translator's initial reading of the fictional characters. Because mind-modelling begins with the reader's self-template, readings of fictional characters will necessarily vary between readers – and translators as readers. Taking into account the time shift between the publication of *Alice* in 1865 and its Czech translation nearly hundred years later, as well as differences between the concept of the child reader in the 19th and 20th century, the translation shifts may also be read as adjustments to the intended audience, e.g. in 20th century Czech, *the Queen* becomes a less terrifying character.

The framework of mind-modelling thus suggests an innovative theorisation of literary translation from which we can derive methodological techniques for the practice of translation. In this chapter, we have illustrated methods for the

comparison of source and target texts. With a focus on the source text, these methods enable translators to identify potentially relevant character cues. The modelling of fictional characters in itself is an unsconcious process, but by identifying potential textual cues explicitly the translator can reflect on choices and consider the potential implications of these choices for the self-templates of the readers of the translation. In this chapter, we drew on reference corpora of children's literature to assess to what extent the translators' choices of character cues for *Alice* and *the Queen* relate to wider patterns. Such a use of reference corpora enables links to the character templates of child readers, as children's literature plays an important role in shaping the experiential qualities of a child's self-template. The theoretical framework we have proposed in this chapter provides a novel approach to literary translation with innovative methodological implications for translation practice, but it also opens up literary translation studies for renewed engagement with literary, linguistic, cognitive and psychological concerns.

Acknowledgements

We would like to thank Michael Toolan and two reviewers for their feedback on previous drafts of this chapter.

References

Ambrosiani, P. 2012. Domestication and foreignization in Russian translations of *Alice's Adventures in Wonderland*. In *Domestication and Foreignization in Translation Studies*, H. Kemppanen, M. Jänis & A. Belikova (eds), 79–100. Berlin: Frank & Timme.

Ben-Ari, N. 1998. The ambivalent case of repetitions in literary translation. Avoiding repetitions: A 'universal'of translation. *Meta* 43(1): 68–78. https://doi.org/10.7202/002054ar

Busse, B. 2010. *Speech, Writing and Thought Presentation in 19th-Century Narrative Fiction*. Habilitationsschrift, University of Bern.

Caldas-Coulthard, D. 1988. *Reporting Interaction in Narrative: A Study of Speech Representation in Written Discourse*. PhD dissertation, University of Birmingham.

Carpenter, H. 2009. *Secret Gardens: A Study of the Golden Age of Children's Literature*. London: Faber & Faber.

Čermáková, A. 2015. Repetition in John Irving's novel *A Widow for One Year*: A corpus stylistics approach to literary translation. *International Journal of Corpus Linguistics* 20(3): 355–377. https://doi.org/10.1075/ijcl.20.3.04cer

Čermáková, A. 2018. Translating children's literature: Some insights from corpus stylistics. *Ilha do Destro A Journal of English Language, Literatures in English and Cultural Studies* 71(1): 117-134. https://doi.org/10.5007/2175-8026.2018v71n1p117

Corness, P. 2009. Shifts in Czech translation of the reporting verb *said* in English fiction. In *InterCorp: Exploring a Multilingual Corpus*, F. Čermák, P. Corness & A. Klégr (eds), 159–176. Praha: NLN.

Fárová, L. 2016. Uvozovací slovesa v překladech třech různých jazyků. In *Jazykové paralely*, A. Čermáková, L. Chlumská & M. Malá (eds), 145–161. Praha: NLN.

Horton, D. 2002. Describing intercultural transfer in literary translation: Alice in Wunderland. In *Kultur und Übersetzung: Methodologische Probleme de Kulturtransfers*, G. Thome, C. Giehl & H. Gerzymish-Arbogast (eds), 95–113. Tübingen: Narr.

Inaki, A. & Okita, T. 2006. A small corpus-based approach to Alice's roles. *Literary and Linguistic Computing* 21(3): 283–294. https://doi.org/10.1093/llc/fqi042

Johnson, J. H. 2016. A comparable comparison? A corpus stylistic analysis of the Italian translation of Julian Barnes' *Il Senso di una Fine* and the original text *The Sense of an Ending*. *Language and Literature* 25(1): 38–53. https://doi.org/10.1177/0963947015623360

Kibbee, D. A. 2003. When children's literature transcends its genre: Translating Alice in Wonderland. *Meta* 48(1–2): 307–321. https://doi.org/10.7202/006977ar

Kruger, A., Wallmach, K. & Murray, J. (eds). 2011. *Corpus-Based Translation Studies. Research and Application*. London: Bloomsbury.

Laviosa, S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

Lathey, G. 2011. The translation of literature for children. In *The Oxford Handbook of Translation Studies*, K. Malmkjær & K. Windle (eds), 198–213. Oxford: Oxford University Press.

Lathey, G. 2015. *Translating Children's Literature*. London: Routledge.

Leech, G. & Short, M. 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Harlow: Pearson Education.

Levý, J. 2011. *The Art of Translation*, transl. by P. Corness [Benjamins Translation Library 97]. Amsterdam: John Benjamins. (The Czech original *Umění překladu* first published in 1963). https://doi.org/10.1075/btl.97

Lindseth, J. A. & Tannenbaum, A. 2015. *Alice in a World of Wonderlands: The Translations of Lewis Carroll's Masterpiece*. New Castle DE: Oak Knoll Press.

Louw, W. E. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 157–174. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.11lou

Louw, B. & Milojkovic, M. 2016. *Corpus Stylistics as Contextual Prosodic Theory and Subtext* [Linguistic Approaches to Literature 23]. Amsterdam: John Benjamins. https://doi.org/10.1075/lal.23

Mahlberg, M. 2005. *English General Nouns: A Corpus Theoretical Approach* [Studies in Corpus Linguistics 20]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.20

Mahlberg, M. 2013. *Corpus Stylistics and Dickens's Fiction*. London: Routledge.

Mahlberg, M., Conklin, K. & Bisson, M-J. 2014. Reading Dickens's characters: Textual patterns and their cognitive reality. *Language and Literature* 23(4): 369–388. https://doi.org/10.1177/0963947014543887

Mahlberg, M. & Smith, C. 2010. Corpus approaches to prose fiction: Civility and body language in Pride and Prejudice. In *Language and Style*, D. McIntyre & B. Busse (eds), 449–467. Houndmills: Palgrave Macmillan. https://doi.org/10.1007/978-1-137-06574-2_26

Mahlberg, M. & Stockwell, P. 2016. Point and CLiC: Teaching literature with corpus stylistic tools. In *Scientific Approaches to Literature in Learning Environments*, M. Burke, O. Fialho & S. Zyngier (eds), 251–267. Amsterdam: John Benjamins. https://doi.org/10.1075/lal.24.13mah

Mastropierro, L. 2017. *Corpus Stylistics in Heart of Darkness and its Italian Translation*. London: Bloomsbury.

Mastropierro, L. & Mahlberg, M. 2017. Key words and translated cohesion – A corpus stylistic analysis of Lovecraft's *At the Mountains of Madness* and its Italian translation. *English Text Construction* 10(1): 78–105. https://doi.org/10.1075/etc.10.1.05mas

Mauranen, A. & Kujamäki, P. (eds). 2004. *Translation Universals: Do They Exist?* [Benjamins Translation Library 48]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.48

McIntyre, D. 2010. Dialogue and characterization in Quentin Tarantino's *Reservoir Dogs*: A corpus stylistic analysis. In *Language and Style*, D. McIntyre & B. Busse (eds), 162–182. Houndmills: Palgrave Macmillan. https://doi.org/10.1007/978-1-137-06574-2_11

Nádvorníková, O. 2017. Les proportions des verbes SAY/DIRE/ŘÍCI dans les propositions incises et leurs équivalents en traduction: Étude sur corpus parallèle. *Linguistica Pragensia* 27(2): 35–57.

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.

Ruano San Segundo, P. 2016. A corpus-stylistic approach to Dickens' use of speech verbs: Beyond mere reporting. *Language and Literature* 25(2): 113–129. https://doi.org/10.1177/0963947016631859

Ruano San Segundo, P. 2017. Corpus methodologies in literary translation studies: An analysis of speech verbs in four Spanish translations of *Hard Times*. *Meta* LXII(1): 94–113. https://doi.org/10.7202/1040468ar

Semino, E. & Short, M. 2004. *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Serbina, T., Hintzen, S., Niemietz, P. & Neumann, S. 2017. Changes of word class during translation – Insights from a combined analysis of corpus, keystroke logging and eye-tracking data. In *Empirical Modelling of Translation and Interpreting*, S. Hansen-Schirra, O. Czulo & S. Hofmann (eds), 177–208. Berlin: Language Science Press.

Shepherd, T. M. G. & Berber Sardinha, T. 2013. *A Rough Guide to Doing Corpus Stylistics*. Matraga, Rio de Janeiro, v. 20 (32), Jan/Jun. <http://www.pgletras.uerj.br/matraga/matraga32/arqs/matraga32a04.pdf>

Siepmann, D. 2015. A corpus-based investigation into key words and key patterns in post-war fiction. *Functions of Language* 22(3): 362–399. https://doi.org/10.1075/fol.22.3.03sie

Sigler, C. (ed.). 1997. *Alternative Alices: Visions and Revisions of Lewis Carroll's Alice Books: An Anthology*. Lexington KY: University Press of Kentucky.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stockwell, P. 2009. *Texture: A Cognitive Aesthetics of Reading*. Edinburgh: Edinburgh University Press

Stockwell, P. & Mahlberg, M. 2015. Mind-modelling with corpus stylistics in *David Copperfield*. *Language and Literature* 24: 129–147. https://doi.org/10.1177/0963947015576168

Teubert, W. 2002. The role of parallel corpora in translation and multilingual lexicography. In *Lexis in Contrast: Corpus-based Approaches* [Studies in Corpus Linguistics 7], B. Altenberg & S. Granger (eds), 189–214. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.7.14teu

Toolan, M. 2009. *Narrative Progression in the Short Story: A Corpus Stylistic Approach* [Linguistic Approaches in Literature 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/lal.6

Toolan, M. 2012. *Poems: Wonderfully Repetitive*. Harlow: Pearson.

Toolan, M. 2016. *Making Sense of Narrative Text: Situation, Repetition, and Picturing in the Reading of Short Stories*. London: Routledge.

Toolan, M. 2018. How children's literature is translated: Suggestions for stylistic research using parallel corpora. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies* 71(1): 151-168. https://doi.org/10.5007/2175-8026.2018v71n1p151

Toury, G. 1977. *Translational Norms and Literary Translation into Hebrew, 1930–1945*. Tel Aviv: The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. 1995. *Descriptive Translation Studies and Beyond* [Benjamins Translation Library 4]. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4

Weaver, W. 1964. *Alice in Many Tongues: The translations of Alice in Wonderland*. Madison WI: University of Wisconsin Press.

Whiteley, S. & Canning, P. 2017. Reader response research in stylistics. *Language and Literature* 26(2): 71–87. https://doi.org/10.1177/0963947017704724

Wittman, E. O. 2013. Literary narrative prose and translation studies. In *The Routledge Handbook of Translation Studies*, C. Millán & F. Bartrina (eds), 438–450. Abingdon: Routledge.

Xiao, R. 2010. How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics* 15(1): 5–35. https://doi.org/10.1075/ijcl.15.1.01xia

Biographical notes

Karin Aijmer
Karin Aijmer is Professor Emerita in English linguistics at the University of Gothenburg, Sweden. Her research interests focus on pragmatics, discourse analysis, modality, corpus linguistics, and contrastive analysis. Her books include *Conversational routines in English: Convention and Creativity* (1996), *English Discourse Particles. Evidence from a Corpus* (2002), *The Semantic Field of Modal Certainty: A Study of Adverbs in English* (with co-author) (2007), *Understanding Pragmatic Markers. A Variational Pragmatic Analysis* (2013). She is co-editor of *Pragmatics of Society* (De Gruyter, 2011), *Handbook of Pragmatics* (Mouton de Gruyter, 2011) and of A *Handbook of Corpus Pragmatics* (Cambridge University Press, 2014), and co-author of *Pragmatics. An Advanced Resource Book for Students* (Routledge, 2012).

Paul Baker
Paul Baker is Professor of English Language at Lancaster University, UK, and a member of the Corpus Approaches to Social Science ESRC Research Centre where he specialises in corpus linguistics and discourse analysis. He has written 16 books, 35 journal articles and 27 book chapters. and is the commissioning editor of *Corpora* journal. His books include *Using Corpora in Discourse Analysis* (Bloomsbury, 2006), *Sexed Texts* (Equinox, 2008), *Discourse Analysis and Media Attitudes* (Cambridge University Press, 2013) (with Tony McEnery and Costas Gabrielatos) and *American and British English: Divided by a Common Language* (Cambridge University Press, 2017).

Michael Barlow
Michael Barlow received his PhD in Linguistics from Stanford University. He is currently Associate Professor in Applied Language Studies at the University of Auckland, New Zealand. Dr. Barlow has written books and articles on corpus linguistics and regularly gives presentations, courses and workshops at institutions and conferences around the world. He has created several text analysis programs including concordancers *MonoConc* and *ParaConc* and a collocation extraction program, *Collocate*. A recently developed program, *WordSkew*, is designed to apply corpus analysis techniques while at the same time taking note of the structure of texts.

Anna Čermáková

Anna Čermáková is currently Marie Curie Fellow at the University of Birmingham, UK, where she works on a project that explores gender in children's literature. She obtained her PhD from Charles University, Prague. She has worked for a number of years at the Institute of the Czech National Corpus, Charles University, where she was involved in the creation of the Czech national corpora. Her research interests are in corpus stylistics, literary translation, contrastive corpus-based linguistics and lexicology. She has written articles in the area of contrastive corpus-based linguistics and translation and published a book with Wolfgang Teubert *Corpus Linguistics: A Short Introduction* (Bloomsbury, 2007).

Sylviane Granger

Sylviane Granger is Professor Emerita at the University of Louvain, Belgium. She is the founder of the Centre for English Corpus Linguistics, of which she was Director for over 25 years. She is one of the pioneers of learner corpus research and has also been active in corpus-based cross-linguistic studies, two fields which she views as closely interrelated. Her current research interests focus on the analysis of phraseology in native and learner language and its integration into electronic writing aids and bilingual dictionaries. Her recent book publications include two volumes on phraseology: *Phraseology: An Interdisciplinary Perspective* and *Phraseology in Foreign Language Learning and Teaching* (with F. Meunier, John Benjamins, 2008), a volume on *Electronic Lexicography* (with M. Paquot, Oxford University Press, 2012) and *The Cambridge Handbook of Learner Corpus Research* (with G. Gilquin and F. Meunier, Cambridge University Press, 2015).

Ramesh Krishnamurthy

Born in Madras (India), Ramesh wrote his first Tamil letter with a stick in a tray of sand at a Hindu ceremony when he was 1, and learned to write with chalk on a slate at infant school. In London, aged 4, he first learned English, using crayons, pencil and paper, and wooden pens with copper nibs dipped in inkwells, then added French, Hindi, Latin, and German (via grammar-translation and close reading) at secondary school. After a degree in French and German at Cambridge, and two years as a programmer and systems analyst in the UK Civil Service, he studied Sanskrit for 7 years at SOAS. He joined the COBUILD project at Birmingham University in 1984, and was immediately fascinated by the use of computers and corpora for lexicography and linguistics. As a corpus linguist, he has created and analysed corpora in many languages, published extensively, taught and supervised undergraduates and postgraduates, lectured and conducted workshops in many countries, and participated in several international research projects. Together with Wolfgang Teubert he is the editor of the six-volume *Corpus Linguistics* published in the Routledge series *Critical Concepts in Linguistics* (2007).

Michaela Mahlberg

Michaela Mahlberg is Professor of corpus linguistics at the University of Birmingham, UK, where she is also the Director of the Centre for Corpus Research and the Director of Research and Knowledge Transfer for the College of Arts and Law. Michaela is the editor of the *International Journal of Corpus Linguistics* (John Benjamins) and together with Wolfgang Teubert she edits the book series *Corpus and Discourse* (Bloomsbury). One of her main areas of research is Dickens's fiction and the socio-cultural context of the 19th century. Her publications include *Corpus Stylistics and Dickens's Fiction* (Routledge, 2013), *English General Nouns: a Corpus Theoretical Approach* (John Benjamins, 2005) and *Text, Discourse and Corpora. Theory and Analysis* (Continuum, 2007, co-authored with Michael Hoey, Michael Stubbs and Wolfgang Teubert). Michaela was the Principal Investigator on the AHRC-funded project *CLiC Dickens: Characterisation in the representation of speech and body language from a corpus linguistic perspective* which led to the development of the CLiC web app.

Alan Partington

Alan Partington is Associate Professor of English Linguistics at Bologna University, Italy. His research interests include corpus research methodology, corpus-assisted discourse study, particularly into political discourses, modern diachronic language studies, evaluation and evaluative prosody, corpus-assisted stylistics, irony, wordplay and metaphor. He is the Editor-in-chief of the *Journal of Corpora and Discourse Studies*, and author of *Patterns and Meanings* (John Benjamins, 1998), *The Linguistics of Political Argument: The Spin-Doctor and the Wolf-Pack at the White House* (Routledge, 2003), *The Language of Persuasion in Politics* (with C. Taylor, Routledge, 2018), *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-Talk* (Routledge, 2006), *Patterns and Meanings in Discourse* (with A. Duguid and C. Taylor, John Benjamins, 2013).

Gill Philip

Gill Philip is Associate Professor of English language and translation at the University of Macerata, Italy, where she teaches undergraduate and masters' courses in corpus linguistics, translation, cognitive linguistics, and TEFL. Most of her research investigates aspects of the interplay between cognition and linguistic communication, particularly with respect to phraseology and figurative language. She has published a number of articles on corpus linguistics and is the author of *Colouring Meaning: Collocation and Connotation in Figurative Language* (John Benjamins, 2011).

Michael Stubbs

Michael Stubbs is Professor of English Linguistics, University of Trier, Germany. He studied at the Universities of Glasgow, Cambridge and Edinburgh, and worked at the Universities of Birmingham, Nottingham and London in the UK before moving to Germany in 1990. He has lectured in Australia and New Zealand, USA and Canada, India, China and Japan, and most western European countries. He has published widely on educational linguistics, text and discourse analysis, and corpus linguistics. He retired from teaching in 2013, but continues to research and write. His current projects are on the computer-assisted analysis of literary texts.

Subject index

With an ever-growing body of corpus linguistic tools, resources and applications, it becomes increasingly important to reflect critically on the underlying assumptions that corpus linguistics is based on. Focusing on meaning and methods, this book tackles fundamental concepts and approaches that define the discourse of the field. Internationally renowned contributors address topics that range from the history of corpus linguistics to contrastive perspectives between languages, to interpreting patterns in corpora as evidence of both mainstream discourses and individual voices within them. This collection not only adds to our understanding of the fundamentals of corpus linguistics, it also brings innovative meanings to the corpus linguistics discourse. It has been edited in honour of Wolfgang Teubert, who for decades has been a significant voice in this discourse.

"This book is a fitting tribute to Wolfgang Teubert – one of the towering figures of contemporary Corpus Linguistics. It takes a critical, reflexive approach to established concepts and shared assumptions in Corpus Linguistics, but also shows what Corpus Linguistics can do, for discourse analysis, contrastive linguistics and translation studies among others. The wide range of topics and methods, and the state-of-the-art approach in each chapter, will appeal to newcomers to the field as well as specialists."

Elena Semino, *Lancaster University*

"This thought-provoking volume brings together an impressive array of contributions by leading researchers in corpus linguistics. Their critical reflections on theory and method do full justice to Wolfgang Teubert's work, and will pave the way for further advances in the field."

Gerlinde Mautner, *Vienna University of Economics and Business*

"In variety and in depth, this volume offers a wealth of discourse. The themes covered within vary widely, but all relate centrally to how we may understand discourse, in all its riches and variety, by analysis using corpus linguistic methods. The volume will provoke lasting reflection and debate and provides timely recognition of the contribution to the field by Wolfgang Teubert."

Mike Scott, *Aston University*

ISBN 978 90 272 0175 1

JOHN BENJAMINS PUBLISHING COMPANY