The volume brings together a selection of invited articles and papers presented at the 4th International CILC Conference held in Jaén, Spain, in March 2012. The chapters describe English using a range of corpora and other resources. There are two parts, one dealing with diachronic research and the other with synchronic research. Both parts investigate several aspects of the English language from various perspectives and illustrate the use of corpora in current research. The structure of the volume allows for the same linguistic aspect to be discussed both from the diachronic and the synchronic point of view. The chapters are also useful examples of corpus use as well as of use of other resources as corpus, specifically dictionaries. They investigate a broad array of issues, mainly using corpora of English as a native language, with a focus on corpus tools and corpus description.

Alejandro Alcaraz-Sintes (Casablanca, Morocco, 1959) is senior lecturer (tenured) at the University of Jaén.

Salvador Valera-Hernández (Jaén, Spain, 1967) is senior lecturer (tenured) at the University of Granada and teaches regularly at Pavol Jozef Šafárik University in Košice, Slovakia.
The volume brings together a selection of invited articles and papers presented at the 4th International CILC Conference held in Jaén, Spain, in March 2012. The chapters describe English using a range of corpora and other resources. There are two parts, one dealing with diachronic research and the other with synchronic research. Both parts investigate several aspects of the English language from various perspectives and illustrate the use of corpora in current research. The structure of the volume allows for the same linguistic aspect to be discussed both from the diachronic and the synchronic point of view. The chapters are also useful examples of corpus use as well as of use of other resources as corpus, specifically dictionaries. They investigate a broad array of issues, mainly using corpora of English as a native language, with a focus on corpus tools and corpus description.

Alejandro Alcaraz-Sintes (Casablanca, Morocco, 1959) is senior lecturer (tenured) at the University of Jaén.

Salvador Valera-Hernández (Jaén, Spain, 1967) is senior lecturer (tenured) at the University of Granada and teaches regularly at Pavol Jozef Šafárik University in Košice, Slovakia.
Diachrony and Synchrony in English Corpus Linguistics
Linguistic Insights

Studies in Language and Communication

Edited by Maurizio Gotti,
University of Bergamo

Volume 181

ADVISORY BOARD

Vijay Bhatia (Hong Kong)
Christopher Candlin (Sydney)
David Crystal (Bangor)
Konrad Ehlich (Berlin / München)
Jan Engberg (Aarhus)
Norman Fairclough (Lancaster)
John Flowerdew (Hong Kong)
Ken Hyland (Hong Kong)
Roger Lass (Cape Town)
Matti Rissanen (Helsinki)
Françoise Salager-Meyer (Mérida, Venezuela)
Srikant Sarangi (Cardiff)
Susan Šarčević (Rijeka)
Lawrence Solan (New York)
Peter M. Tiersma (Los Angeles)
Alejandro Alcaraz-Sintes & Salvador Valera-Hernández (eds)

Diachrony and Synchrony in English Corpus Linguistics
Contents

Acknowledgements...................................................................................................................... 7

BAS AARTS
Preface........................................................................................................................................ 9

Part 1: Corpora and Historical Linguistics

ALEJANDRO ALCARAZ-SINTES
Dictionary- and corpus-based research in historical linguistics........ 13

NURIA CALVO-CORTÉS
A corpus-based study of gradual meaning change in Late Modern English.............................................................. 25

TERESA FANEGO
Dictionary-based corpus linguistics and beyond: developments in the expression of motion events in the history of English........ 55

MARÍA JOSÉ LÓPEZ-COUSO / BELÉN MÉNDEZ-NAYA
The use of if as a declarative complementizer in English: theoretical and empirical considerations....................................... 85

MATTI RISSANEN
On English historical corpora, with notes on the development of adverbial connectives......................................................... 109

ONDŘEJ TICHÝ / JAN ČERMÁK
Measuring typological syntheticity of English diachronically with the use of corpora.............................................................. 141
Part 2: Corpora and Descriptive Linguistics

SALVADOR VALERA-HERNÁNDEZ
Dictionary- and corpus-based research in applied and descriptive linguistics .......................................................... 161

MIGUEL-ÁNGEL BENÍTEZ-CASTRO
Formal, syntactic, semantic and textual features of English shell nouns: a manual corpus-driven approach ................... 171

EDUARDO COTO-VILLALIBRE
From prototypical to peripheral: the ‘get + Ven’ construction in contemporary spoken British English .......................... 205

THOMAS EGAN
Encoding ‘throughness’ in English and French ........................................ 233

BEATRIZ MATO-MÍGUEZ
If you would like to lead: on the grammatical status of directive isolated if-clauses in spoken British English ............ 259

DETMAR MEURERS / JULIA KRIVANEK / SERHIY BYKH
On the automatic analysis of learner corpora. Native Language Identification as experimental testbed of language modeling between surface features and linguistic abstraction ................................. 285

JUAN SANTANA-LARIO
‘Adjective + whether/if-clause’ constructions in English. An exploratory corpus-based study ........................................ 315

PAUL THOMPSON
Exploring Hoey’s notion of textual colligation in a corpus of student writing .......................................................... 347

Notes on contributors .................................................................. 373

Subject index .................................................................................. 381
Acknowledgements

We gratefully acknowledge the financial support of the Spanish Ministry of Economy and Competitiveness (grant FFI2011-15186-E), which has helped cover the organizational costs of the 4th International Conference on Corpus Linguistics organized at the University of Jaén (Spain) (22nd-24th March 2011) on behalf of AELINCO (Spanish Association for Corpus Linguistics) by A. Alcaraz-Sintes, M. A. Benítez-Castro, A. V. Casas-Pedrosa, A. Díaz-Negrillo, F. J. Díaz-Pérez, J. Fernández-Domínguez and S. Valera-Hernández, as well as the publication of this book containing selected papers and guest contributions.

We are also grateful for the assistance given by a number of colleagues who acted as peer-reviewers in the selection process.

The Editors
Jaén, June 2013
Preface

The scholarly articles in this volume are based on presentations delivered at the Fourth International Conference on Corpus Linguistics (CILC) held at the University of Jaén from 22 to 24 March 2012 under the aegis of AELINCO, the Spanish Association of Corpus Linguistics. The conference attracted almost two hundred scholars from a large number of countries in several continents.

The book before you contains a selection of the papers presented at the conference and is organised into two parts: one focuses on diachronic, the other on synchronic research. Within these two parts the various chapters cover a very wide span of research topics, mostly on English, and each of them attests to the lively issues and debates currently taking place in the burgeoning field of Corpus Linguistics.

This book is a stimulating read and is certain to make an important and lasting contribution to our knowledge.

Bas Aarts
University College London
June 2013
Part 1

Corpora and Historical Linguistics
Dictionary- and corpus-based research in historical linguistics

Abstract

Research on historical linguistics conducted with corpora has for quite some time now come to be accepted as the most reliable type of research, not only to confirm or hypothesize on patterns and causes of linguistic change, but also to test different theoretical models applied to the study of specific grammatical issues and to perform accurate descriptions of language features in or through past synchronies. The present chapter analyses the five contributions in Part 1 insofar as they bear witness to this claim. Five very specific grammar issues are addressed and investigated: the Late Modern English semantic changes and grammaticalization processes undergone by a closed set of nautical $a$-prefixed words originally denoting location and direction, the evolution of verbs expressing manner of the motion, the development of the adverbial subordinator $if$ as a declarative complementizer down to Modern English, the progress of three specific adverbial connectives along the history of English as evidenced by data gathered from different historical corpora, and the manner in which the evolution of English over the centuries as regards its analicity versus syntheticity typology can be measured and established.

1. Introduction

One of the strengths of corpus-based research in historical linguistics lies in the increasing availability of historical corpora, large and small,
Alejandro Alcaraz-Sintes

(see CorD),² in the wider range of corpus tools that can be used on corpora and, as a result of both, in the relative ease with which data and results can be verified or falsified by the researchers.

The corpora used by the contributors also reflect the trend in corpus compilation from the small-sized “long-diachrony” or “multi-purpose” (Rissanen 2000: 8) corpora, such as the Helsinki Corpus HC³ or ARCHER,⁴ towards larger but period- or genre-specific ones, such as the CEEC⁵ or the CED,⁶ to name just a couple. In other words, the overall trends for certain phenomena observed in data obtained from general corpora are checked against those from “short and fat” (Kohnen 2009: §2.1) ones. At the same time, some contributors have also obtained their data from dictionary quotation databases, mainly that of the OED,⁷ and full-text collections (equivalent for certain purposes to genuine corpora; see Hoffman 2004; Mair 2004: §2; Lindquist 2009: §9.4).

There are differences between the contributions as regards the descriptive, theoretical and explanatory approaches taken by their authors to describe or explain the phenomena studied. As Rissanen (this volume) aptly writes in his introduction, “[t]heoretical considerations add to the value of data-based results, while surveys of data are essential to trigger and support theoretical discussions.”

2. Studies

Nuria Calvo-Cortés’ detailed study on how four a-words underwent metaphoric extensions through different processes of grammaticalization confirms a couple of requisites to relate the full story of specific

---

³ Helsinki Corpus (Rissanen/Kytö 1991).
⁴ A Representative Corpus of Historical English Registers (ARCHER 1993).
⁵ Corpus of Early English Correspondence (1998).
⁶ Corpus of English Dialogues (Kytö/Culpeper 2006).
⁷ Oxford English Dictionary (Simpson 2000).
grammatical or lexical problems. On the one hand, the importance of the ModE period for many changes (also manifest in other contributions), particularly LModE, and, on the other hand, how the data in the OED can be complemented by corpus-based data. The author has used the 15-million-word CLMETEV.8 Through the detailed study of examples found in the corpus she identifies meanings hitherto unrecorded by the OED and presents data for each period and word, specifying with accuracy when the semantic changes took place.

Teresa Fanego’s contribution on “verbs of sound emission” shows that historical dictionaries, thesauri and literature collections too, namely, the OED, the MED,9 the HTOED10 and the Chadwyck-Healey Literature Collections Online, offer sufficiently large amounts of data to trace the history of words and grammatical constructions. Her particular object of study – verbs that combine the meanings of motion and manner of motion – are approached from the perspectives of cognitive semantics and Constructional Grammar (Goldberg 1995, 2006; Talmy 2000; Slobin 2004a, 2004b, 2006). While Fanego’s previous studies on motion verbs (e.g., Fanego 2012) show that the stock of such verbs in the English lexicon has been constantly added to in the different periods of English, in this paper she focuses specifically on “verbs of sound emission” (where motion and manner generate a sound) whose appearance in the Early Modern English (EModE) period was made possible or “motivated” (cf. Lakoff 1987) by the previous existence of manner verbs. The new intransitive “sound emission to motion” construction has gradually spread in usage and members since then, not only for reasons of “motivation”, but also of phonesthesia (many verbs ending in /ʃ/), and the construction has in fact become firm enough now to integrate verbs other than sound verbs.

Unlike the previous two authors, María José López-Couso and Belén Méndez-Naya’s approach to their object of study, the long history of the conditional link if used as a declarative complementizer from OE to Present-Day English (PDE), is based on data obtained

8 Corpus of Late Modern English Texts (Extended Version) (De Smet 2006).
9 Middle English Dictionary (Kurath et al. 1952-2001).
10 Historical Thesaurus of the Oxford English Dictionary (Kay et al. 2009).
from two general corpora, the *HC* and *ARCHER*. Interestingly, this type of structures overlap in part with those studied by Santana-Lario in the second part of this volume. After a revision of the established criteria to be met by this *if*-construction to be considered a complement equivalent to a *that*-construction, López-Couso and Méndez-Naya present a thorough synchronic and diachronic description of the data obtained from the corpora and the trends observed for the different periods. One of the most interesting results is reached thanks to the filtering capacities of *ARCHER* for registers: *if*-complements are more commonly found in the speech registers than in the written ones, and in the less formal text-types in either register. This confirms precisely what the authors have found to be the case for other minor complementizers in previous research.

Matti Rissanen’s paper seeks to show what historical corpora can be used to study linguistic change and, what is more important, how. For this purpose he looks into the behaviour and diachronic evolution of three adverbial connectives. Like other contributors, Rissanen does not merely offer a detailed review of their evolution, but also provides explanations from the perspectives of grammaticalization and language contact. For the OE period, the author analyzes the connective *nemne/nympe* ‘except’ and, at the same time, illustrates the differences between two corpora, the all-inclusive *DOEC*¹¹ and the TEI XML version of the *HC*,¹² which permits searching by date, dialects and genres. While the data retrieved from the latter speedily show that the term may be considered as non-West Saxon and poetical, the *DOEC*, though yielding all the occurrences of the word, is less amenable to searching.

For the Middle English (ME) and later periods Rissanen focuses on a couple of French-based connectives. The first one, *according to*, started to be used as a preposition in the 15th century and soon became grammaticalized, three centuries after the verb *accord* had been borrowed. By comparing the relatively scant data from the ME part of the *HC* and the more abundant ones from the larger

---

ICoMEP\textsuperscript{13} and CMEPV,\textsuperscript{14} Rissanen proves that grammaticalization from verb to preposition started in the 15th century and was well established in the 16th century. Equally important, the \textit{HC} data reveal that most ME tokens are found in statutory texts. Similar trends are manifest in data for the EModE part of the \textit{HC} and in the PPCEME.\textsuperscript{15} These findings justify – Rissanen rightly argues – the inclusion of this genre in any Late Middle English (LME) corpora. Unlike \textit{according to}, the second connective, \textit{concerning}, was introduced at the same time as the verb in very late ME as an already grammaticalized form and became firmly established in the 17th century, according to \textit{HC}, ICoMEP and CMEPV data. Besides, on the evidence from the \textit{HC}, although \textit{concerning} was used in many different text-types, it is particularly present in formal registers. The availability of genre-specific EModE corpora, such as the \textit{PCEEC},\textsuperscript{16} the \textit{CEEM},\textsuperscript{17} and the \textit{CED}, does not only throw light on the speed at which the form spread from formal to other registers, but also confirms the increasing rate of occurrence of \textit{concerning} (and \textit{concern}) in 17th century dialogue texts from trials and depositions, and in medical texts. The evidence gathered by the author for LModE and the 20th century from \textit{ARCHER} and much larger corpora, such as the \textit{CLMETEV}, the \textit{PPCMBE},\textsuperscript{18} the \textit{COHA},\textsuperscript{19} \textit{Brown},\textsuperscript{20} \textit{Frown},\textsuperscript{21} the \textit{LOB},\textsuperscript{22} the \textit{F-LOB},\textsuperscript{23} the \textit{BNC}\textsuperscript{24} and the \textit{COCA},\textsuperscript{25} showing that connective \textit{concerning} dramatically de-

\textsuperscript{13} Innsbruck Corpus of Middle English Prose (Markus 2010).
\textsuperscript{14} Corpus of Middle English Prose and Verse (Humanities Text Initiative, University of Michigan 2006).
\textsuperscript{15} Penn-Helsinki Parsed Corpus of Early Modern English (Kroch et al. 2004).
\textsuperscript{16} Parsed Corpus of Early English Correspondence (Taylor et al. 2006).
\textsuperscript{17} Corpus of Early English Medical Writing (Taavitsainen/Päivi 2010).
\textsuperscript{18} Penn Parsed Corpus of Modern British English (Kroch et al. 2010).
\textsuperscript{19} Corpus of Historical American English (Davies 2010-).
\textsuperscript{20} The Standard Corpus of Present-Day Edited American English (Francis/ Kučera 1964).
\textsuperscript{21} The Freiburg-Brown Corpus of American English (Mair 1999, 2007).
\textsuperscript{22} The Lancaster-Oslo/Bergen Corpus (Leech/Johansson (1976, 1986).
\textsuperscript{23} The Freiburg-Lancaster-Oslo/Bergen Corpus (Mair/Leech 2007).
\textsuperscript{24} The British National Corpus (2007).
\textsuperscript{25} The Corpus of Contemporary American English (Davies 2008-).
creases in usage, unlike verbal forms of concern, proves that some kind of register-based restriction seems be in force, hindering its complete grammaticalization. Rissanen’s chapter, therefore, not only shows that using general corpora permits to obtain overall frequencies of items or infer trends of linguistic change, but that descriptions of language change will be more fine-grained and truer to actual language use and synchronic variation if they rely on corpora that allow filtering results by text-genre, dialect and sociolinguistic factors.

Finally, Tichy and Cermak’s contribution to this volume deals with the evolution of English as regards its typological status, analytical or synthetic, and attempts to test whether this status is susceptible of measurement by conducting a series of morphological and statistical analyses on historical corpora (The York-Toronto-Helsinki Parsed Corpus of Old English Poetry (Taylor et al. 2001), the YCOE\textsuperscript{26} and the BNC). Following Szmrecsanyi (forthcoming), the authors define analycity as the frequency of free grammatical markers, and syntheticity as the frequency of bound grammatical markers. According to Szmerecanyi (forthcoming), the change of English towards analycity started to decline in the EModE period, that is, syntheticity was on the rise. This is directly contrary to the accepted view on the typological history of English. However, the authors’ refined measurement of the trends in the different periods of English (though excluding different varieties of English and measurements of analycity) and the application of statistical tests to verify their significance confirm, contrary to Szmerecanyi’s claim, the syntheticity-to-analycity evolution of English. The main contribution is to show how to use corpora, particularly tagged corpora, in order to measure the frequency of distribution of morphological markers across paradigms and how an index of syntheticity for a given word-class and period can be given by means of a formula. The authors also show that different word-classes differ in their syntheticity levels (verbs and adjectives, become less homogeneous, but nouns do not) and that, within the same class, some words may lose their syntheticity, while others may increase it, as exceptional residues of an older system.

\textsuperscript{26} The York-Toronto-Helsinki Parsed Corpus of Old English Prose (Taylor et al. 2003).
References

ARCHER = A Representative Corpus of Historical English Registers. 1993-. <manchester.ac.uk/archer>.


Corpus of Early English Correspondence. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of Modern Languages, University of Helsinki.


Fanego, Teresa 2012. Motion Events in English: The Emergence and Diachrony of Manner Salience from Old English to Late Modern English. Folia Linguistica Historica. 33, 29-85.


NURIA CALVO-CORTÉS

A corpus-based study of gradual meaning change in Late Modern English¹

Abstract

The present study focuses on the analysis of a set of originally nautical terms (aboard, ahead, aloof and astern) in the Late Modern English period. Specifically, the main interest lies in showing the contribution of corpora to the analysis of meaning change. The examples containing the chosen terms were extracted from the Corpus of Late Modern English Texts (Extended Version) (CLMETEV). They were all used for the first time on board ships, i.e., they were part of the nautical jargon, and their meanings were originally connected to the notions of location and/or direction. However, most of them extended their uses to other contexts because of different processes of grammaticalization and developed metaphorical extensions. The analysis of the examples as well as the contexts in which they appeared was followed by a series of statistical analyses to compare the evolution of the meanings and the frequency of usage of the different words. The results show that, in general, whereas at the beginning the literal original meanings of the words were more frequently present, towards the end of the period they had already achieved the range of meanings that can be found in the Oxford English Dictionary and even others not present in it. Also, their presence was restricted to very specific texts, mainly connected to the world of the sea, in the early 18th century, whereas later on in the period most of the texts included in the corpus contained at least one of the terms. Finally, it can also be observed that towards the end of the period some collocations between the chosen terms and some verbs become gradually closer and, therefore, more grammaticalized.

¹ Author’s email address: ncalvo@filol.ucm.es; affiliation: Complutense U. of Madrid.
1. Introduction

The *Oxford English Dictionary (OED)* (Simpson/Weiner 1992) provides us with the examples showing the first written instances of the different meanings of words in the English language. As the language is constantly changing, the dictionary is gradually updated to show these changes. However, when analysing meaning change, because the updating process can be slow, corpora can provide a clearer insight into the reasons for these changes as well as into the contexts where the new meanings first appeared. Therefore, a combination of both, the *OED* and corpora, will prove essential in any study of English language meaning change.

The use of corpora to analyse language evolution and meaning change has been common practice in the last few decades (Facchinetti/Rissanen 2006; Nevalainen et al. 2008; Claridge 2010). However, the Late Modern English (LModE) period (18th and 19th centuries) has been slightly neglected in historical linguistics if compared to the previous periods of the language. As a consequence, studies concerning the language of that period (Beal 2004; Tieken Boon van Ostade 2009; Barber et al. 2009), as well as the production of corpora (*A Corpus of Late 18th Century Prose* (Denison/van Bergen 2003); *The Corpus of Late Modern English Texts (Extended Version)* (CLMETEV) (De Smet 2006)) containing texts dating from that time have only gained interest in the last decade. Although it is a period in which many grammatical aspects were already very similar to those of Present Day English (PDE), in the LModE period many words developed new meanings and there was a growth in vocabulary thanks to several factors external to the language itself, such as the Industrial Revolution.

The aim of the present study is to analyse the changes of meaning experienced by four originally nautical terms, namely, *aboard, ahead, aloof* and *astern*, in the LModE period, more precisely, to focus on the contribution of corpora to the analysis of meaning change, as opposed to the use of the *OED* alone.
The four terms chosen for the purpose of this analysis share some grammatical features. They all contain the prefix *a-* (/ə/), which derives from the Old English (OE) preposition *on*, and a noun related to one physical part of a ship (*board, head, stern*) or something in its immediate physical space (*loof*). In addition, the four prefixed forms are present in similar syntactic contexts, that is, they very often function as locative complements accompanying verbs of either motion or location, such as *go* or *be*.

Since these terms were all used for the first time on board ships, i.e., they were part of the nautical jargon, their meanings were originally connected to location or movement in relation to a ship. They all developed metaphorical extensions and extended their uses to other contexts because of different processes of grammaticalization. The new, more abstract meanings began to appear in the 18th century, with the exception of *aloof*, which had already developed metaphorical meanings in the Early Modern English (EModE) period (which extends from the end of the 15th century to the end of the 17th century).

Even though the set of words analysed in this study is very limited, it shows some patterns of usage and meaning change that can be applied to other related words that may share some grammatical features with them. They include terms such as *aback, afore, athwart, before*, which also contain a prefix, either *a-* or *be-*-, and have undergone similar processes of language change that have resulted in meaning changes and/or meaning extensions (e.g., *aback* originally had a physical sense in relation to “motion backwards” and in the 19th century it began to be used to mean ‘surprised’, particularly in the expression *taken aback*, as it developed a new figurative meaning (*OED*)).

1.1. The original meanings and the new meanings

Giving the date of the first time a word has ever been used is almost an impossible task, since it may have been used orally for a long time before it was actually written in a non-private text such as a novel. This is one of the reasons why giving the first attested usage of a word
is not easy. However, the *OED* attempts to do this by providing examples of any word being used as early as records have been found.

Stating the original meaning of any given word is not always easy either. This is particularly evident when a word is polysemous and its origin is not clear. However, in most cases an etymological analysis of a word usually contributes to the clarification of its original more basic meaning. The etymological orientation of the *OED* makes it an invaluable source for the discovery of the primary meaning of any word.

The *OED*’s definitions of the original meanings of the four selected terms, as well as the date of the first example recorded with those meanings, are as follows:

- *Aboard* (1494): “on one side (of a ship\(^2\) or shore)”.  
- *Ahead* (1596): “in a position or direction pointing forward”.  
- *Aloof* (1532): “away, to the windward from within a ship”.  
- *Astern* (1627): “in the rear, behind (at any distance)”.  

As can be observed, they all refer to either location or motion in relation to a ship. Although apparently clear, these definitions are slightly vague, as they refer to a very wide and not very well defined location.

These four words developed some new meanings, most of which are included in the *OED*. The new meanings relevant for this study as well as the dates of the first examples of them provided by the *OED* are as follows:

---

2 Where *board* refers to the planks on the sides of the ship.
3 From within a ship.
4 In relation to the *stern* of a ship.
5 Some of the meanings found in some examples in the corpus were not present in the *OED*, as will be pointed out accordingly.
6 The meanings that are relevant for this study are those found in the examples extracted from the corpus, as will be analysed below.
7 *Astern* is not included in this list because the *OED* does not make it clear whether any of the meanings might imply the use of *astern* in relation to other means of transport apart from ships, as will be explained below.
A closer analysis of the meanings of the nouns forming these words will help to clarify these original meanings and contribute to a better understanding of why they have undergone the specific metaphorical extensions that were already present in LModE.

1.2. The meanings of the nouns board, head, loof and stern

As the definitions of the four prefixed words offer some ambiguity, it is necessary to analyse the meanings of the nouns to which a- was prefixed. At first sight, these nouns appear to be complex. Board and head are polysemous, whereas loof and stern have other forms of complexity. These features will be essential to understand the evolution of the metaphorical meanings of the prefixed forms.

Board is not a term used exclusively in the nautical jargon. Its meaning is broad, as it may refer to any type of plank. This means that, apart from being used in relation to the planks that form the deck or the hull of a ship, it is also used in relation to the top part of a table or any other object containing a long flat piece of wood. When the combination aboard appeared for the first time, according to the definition given by the OED, it seems that the noun board referred to the planks forming the hull rather than the ones forming the deck. However, it is arguable that this meaning of board can be challenged, since being on board (aboard) could literally refer to being physically located “on the boards”, that is, on the planks forming the deck. If this was the case, it might be easier to understand the extension of aboard
to other means of transport, in which one is located definitely on something and not necessarily surrounded by any structure. Furthermore, some of the more metaphorical expressions of *aboard*, such as *have too much drink aboard* or *take something aboard* (or *on board*), may have also been influenced by the meaning of *board* in relation to the top of the table.

Similarly, *head* is not a maritime term in itself. Although the combination *ahead* was first used on board ships, the noun *head* was already being used metaphorically in this prefixed form because it referred to the front part of the ship, by extension of the body part. There are two possible explanations for this. First, many of the large ships in the 16th century had a figure-head located in the very front part of the ship; secondly, the ship resembled a fish opening the sea with its *head*, its front part (Smyth 1867: 374). Both interpretations indicate that the *head* actually referred to something that was located in the external front part of the ship, which explains the fact that, when indicating that something is *ahead* in relation to somebody located inside a ship, it is never inside, but outside the ship. The more abstract and metaphorical meanings that *ahead* developed later always refer to something external as well.

The term *loof* is more complicated to analyse. It does not refer to a specific position either in relation to a ship or to anything else; it is used for a much less definite area, since it refers to “the place where the wind comes from”, which is a variable location. In addition, the etymology of *loof* (*luff*) is complex. The *OED* states that its origin is uncertain and might have been incorporated into English from either Dutch or French, and its original meaning is not clear either. These factors have probably contributed to the meaning of the term *aloof*, which does not have a fixed location in relation to a ship, but depends on the area from which the wind comes. As regards the more metaphorical meanings of *aloof*, they can easily be understood when the original meaning is interpreted as ‘away’, which clearly points at being at a distance.

The meaning of *astern* also needs to be clarified by analysing the meaning of the noun *stern*. The etymological information provided by the *OED* is very clarifying. If it is considered that the original meaning of *stern* probably derived from the Old Norse form for the
tiller, the situation is very similar to the one explained for *ahead*, i.e., the original meaning of *astern* would be connected with something located outside the ship, and never inside. However, as opposed to *ahead*, *astern* did refer to location inside the ship as soon as the word appeared, because the meaning of *stern* had already been extended to refer to a wider area, which included the inside part of the rear of the ship. This explains the definition given by the *OED* for *astern* (s.v. *astern*, A. adv. 2): “in the rear, behind (at any distance)”, where the preposition “in” is used.

1.3. Processes involved in the appearance of the new meanings

It is clear that the meaning of the nouns joining the prefix *a*- had a very powerful influence on the development of all the meanings of the new prefixed forms, both the original literal ones and the more abstract and metaphorical ones.

However, the appearance of the new prefixed terms, as well as the semantic extensions, involved several typical processes of language change, which include lexicalization and grammaticalization.

Although both lexicalization and grammaticalization arise in “the spontaneous and productive combination of lexical items in discourse” (Himmelmann 2004: 36), they refer to different phenomena. However, according to Lehmann (2003: 17), both imply reduction and they can “occur jointly in a given case” (Lehmann 2003: 18), with lexicalization taking place first, as Rostila (2004: 1) also indicates.

The terms analysed in this paper underwent a lexicalization process before becoming more grammatical elements, and, therefore, before being used in different syntactic structures with different meanings.

The process of lexicalization that these terms underwent should be understood as “the change whereby in certain linguistic contexts speakers use a syntactic construction or word formation as a new contentful form with formal and semantic properties that are not completely derivable or predictable from the constituents of the
construction or the word formation pattern” (Brinton/Traugott 2005: 144). It is important to emphasize that it is not only one element that lexicalizes, but a whole structure, which is reinterpreted as a new lexical unit.

This process took place at a very early stage, when the prepositional phrases “on board”, “on head”, “on loof” and “on stern” became new lexical items and started appearing in the early dictionaries under their own entries (e.g., Barlow 1772 and Ash 1795). The original OE preposition on (later reduced phonologically to /ə/) joined the noun that was functioning as its complement. Therefore, the syntactic construction, that is, the prepositional phrase, was not interpreted analytically, but holistically (Lehmann 2003: 2), as a whole new unit, which gradually began to be used in other contexts. The new words were originally considered adverbs, but they soon started to be followed by complements, which means that they could then be analysed as prepositions (following the traditional distinction established between adverbs and prepositions, as explained below). These extensions to other categories of words together with the development of more abstract meanings are an indication that a process of grammaticalization was already happening in the LModE period.

Brinton/Traugott (2005: 145) define grammaticalization as “the change whereby in certain linguistic contexts speakers use parts of a construction with a grammatical function”. This process is, therefore, pragmatically driven and according to Lehmann (in Roberts/Roussou 1999: 1012) it involves attrition, that is, the element loses semantic and phonological content; condensation, which explains why the new forms become simpler in terms of the complementation they can take; paradigmatization, that is, the reduction of the grammaticized forms to morphology-like material; coalescence, which implies that often free morphemes become bound morphemes; obligatorification, that is, the new grammaticalized forms are required by the syntactic structure, unlike lexical items; and, finally, fixation, which means that elements that undergo grammaticalization usually appear in fixed positions either in the syntactic or morphological structure.
The four words analysed here experienced loss of semantic content. For example, the meaning of *aboard* in *she stepped aboard*\(^8\) is semantically richer than the meaning of *aboard* in *all aboard the train*, since in the former example *aboard* still maintains the original meaning of *board*, either “the planks that formed the deck of the boat” or “those surrounding the boat forming the borders of it”, whereas in the second example it has lost that original meaning. The type of complementation they can take is also very limited, for instance, *aloof* is always followed by the preposition *from* when it has a complement, as in *he stood aloof from us*. They also become obligatory elements in the syntactic structure of statements indicating motion situations (e.g., *he came aboard*) or more metaphorical expressions (e.g., *they were ahead of us*, where *ahead* is referring to time), where they have a fixed position, after the verb.

The category of the prepositions has traditionally been considered much more grammatical than the category of adverbs, due to semantic and syntactic reasons. Some prepositions hardly contain lexical meaning (e.g., *by, of, for, to*) and their main function is syntactic (van Gelderen 2002: 19), that is, they serve to introduce different types of complements in the sentence (e.g., *the book was written by Smith, the leg of the table, these flowers are for you, he gave the letter to me*). In addition, prepositions have always been considered to require a complement (*Smith, the table, you, me*). On the contrary, adverbs are usually richer in semantic content (e.g., *beautifully, clumsily, fast*) and they do not require a complement, unlike prepositions. However, the prepositions analysed in this paper have much more lexical content, which is even more evident when they are used without a complement (e.g., *he climbed aboard, they went ahead*). As Déchaine (2005: 4) points out regarding the “status of prepositions with respect to the open-class/closed-class distinction”, prepositions belong to a “borderline category” since they share features of both word-classes.

The obligatoriness of the presence of a complement, as has been pointed out, has been a traditional argument to distinguish adverbs

\(^8\) All examples not including a bibliographical reference have been made up by the author to clarify different aspects.
from prepositions (Burton-Roberts 1992: 167). This was already present in grammars of the 18th century, such as Lowth (1762: 6, 64) and was maintained in 19th century grammars (Latham 1855: 321; Adams 1858: 78). Also, manuals published in the 20th century clearly stated that prepositions had little lexical content and had a connective function (Funk and Wagnalls 1953: 3; Whitehall 1958: 53) As opposed to that, the present-day distinction between transitive and intransitive uses of prepositions (Radford 1997: 269; Lee 1999: 136; Huddleston/ Pullum et al. 2002: chapter 7) can be applied to these four terms. They function intransitively when they are not followed by any complement, as in *he went aboard*, whereas their function is transitive when a complement stands next to them, as in *he went aboard the boat*. The four prepositions show differences in relation to the type of complement following them. While *ahead* and *astern* are complemented by a prepositional phrase introduced by *of* (*ahead of* or *astern of us*), *aloof* is also followed by a prepositional phrase, but introduced by *from* (*aloof from them*), and *aboard* takes a noun phrase as its complement, as in the example just mentioned, *he went aboard the boat*. These differences are part of another ongoing survey and they go beyond the scope of the present study; for this reason they will not be analysed here.

The new meanings and the new syntactic functions did not develop very quickly. This was mainly due to the fact that the prescriptive grammarians in the 18th century condemned the usage of these *a*-prefixed words because they considered them the result of “familiar” and “not very careful” speech, as Lowth (1762: 65) clearly stated in his grammar. On the contrary, in the 19th century the presence of these forms boomed9 as the grammar books of that time did not follow the condemnatory practices that the previous grammarians had expressed. It is also the century in which there

9 The present study is part of a larger one including more maritime terms that support this tendency. As for other *a*-prefixed words, a study by Newmann (1943) confirmed the increase in the number of these words in the 19th century. He also indicated that “some of the words which appear for the first time in the nineteenth century literature may have existed in colloquial or dialectal speech earlier, but were unrecorded in literary documents” (Newmann 1943: 281) due to the influence of 18th century grammarians.
seems to have been a great interest in these prefixed forms in general, as shown by the amount of articles written specifically on these words (see, among others, Piltz 1851; Regel 1855; Skeat 1874). Figures 1, 2, 3 and 4 confirm the growing usage of the *a*- prefixed words in the 19th century.

Despite sharing a similar morphological evolution, the four words did not develop the same metaphorical extensions. However, all of them clearly extended their meanings beyond the physical locative or directional senses they originally had. In addition, whereas *aloof* acquired an abstract meaning at a very early stage, almost as soon as the new term was created, *aboard* and *ahead* took longer to extend their usage to other contexts. *A stern* proves to be more limited in its development because it is hardly ever used outside the maritime contexts. The different meanings and the development of the metaphorical extensions will be analysed further in §3.

The definition of *metaphor* varies according to different authors. In the present paper *metaphor* is understood as “the use of language to refer to something other than what it was originally applied to, or what it ‘literally’ means, in order to suggest some resemblance or make a connection between the two things” (Knowles/Moon 2006: 3). This is a very broad definition, since it encompasses almost any change of meaning within a given word, that is, processes such as metonymy and generalization could be included under this definition as well. While all the terms chosen seem to have undergone a similar process of lexicalization and grammaticalization, not all of them have experienced great variation in terms of meaning. Therefore, this definition proves appropriate for the analysis of the meaning changes of the four words.

As will be explained in §3, the terms that seem to have developed more abstract and metaphorical meanings are *ahead* and *aloof*, whereas *aboard* and *astern* extended their original meanings to places outside the sea, but did not acquire very figurative meanings (with the exception of *aboard* in *having drink aboard*; see §3.1). One of the most significant of these abstract meanings, shown by *ahead*, is the “space-to-time extension”, which is a clear example of a metaphor-

---

10 As could be seen in §1.2. and will be further explained in §3.
atical extension that has been repeatedly described in the literature (among others, Lakoff/Johnson 1980; Kövecses 2002; Knowles/Moon 2006).

Finally, towards the end of the LModE period, new processes of lexicalization can be observed, since there are new combinations of lexical items that seem to behave as one lexical unit, as is the case of all aboard and go(-)ahead. The former seems to be frequently used in connection to trains, whereas the second one does not seem to be fully established in the LModE period insofar as it shows variation in form, which includes going-ahead and go ahead. In these expressions the a- prefixed words seem to be the ones that carry the semantic meaning. In the case of all aboard there appears to be a verb missing (come) and it seems as if aboard is actually taking on the function of both the verb and its complement. As for go(-)ahead, despite the fact that there is a lexical verb in the structure (go), this also seems to have lost most of its lexical content in favour of the content of the preposition ahead. They are both frequent expressions in Present-day English and go(-)ahead is even becoming more and more grammatical, since it is extended to other categories of words, such as nouns, as evidenced by examples from the British National Corpus (BNC) (e.g., “Finally I was given the go ahead and everything was resealed” – CLT 1444; “now we’ve got the go ahead for the new chiller” – JTC 150). However, this will not be dealt with in the present paper as it goes beyond the scope of this topic.

2. Methodology

The four terms selected for the present study were analysed both indirectly, in the definitions provided by the OED, which contributed to the clarification of meanings in general, and directly, in the examples extracted from the Corpus of Late Modern English Texts (Extended Version) (CLMETEV) (De Smet 2006).
The OED was used to check, first of all, all the different meanings that the chosen terms had and, secondly, to observe the approximate period in which these different meanings may have appeared for the first time. The dates showed that two of the chosen terms, aboard and ahead, had developed metaphorical extensions in the LModE period (in the OED, the first example of aboard used in relation to another means of transport is from 1855, whereas the first example of ahead used with a temporal meaning is from 1900). They also showed that aloof had acquired more abstract meanings in the previous period, i.e., the EModE period (the first example of aloof used in the sense of distance in relation to feelings is from 1583), whereas in the entry for astern no metaphorical meanings were present at all.

The LModE period coincided with the Industrial Revolution, a time in which ships were already developing new means of power of propulsion, such as steam, with a consequent decline in the use of sail power, resulting in a decrease in the evolution of the terminology directly related to wind in the nautical jargon.

For all these reasons, we considered it more appropriate to choose a LModE corpus to analyse these terms. The CLMETEV includes some specific maritime works, such as Captain Cook’s Journal (1768-1771) or The Life of Horatio Lord Nelson (Southey 1813), but it also includes many novels and essays on a great variety of topics. The contents of the corpus together with the fact that the total number of words amounts up to almost 15 million contributed to its choice as opposed to other corpora of the period (e.g., A Corpus of Late Modern English Prose, which contains 100,000 words; A Corpus of Late 18th Century Prose (Denison/van Bergen 2003), which includes 300,000 words for the period 1761-1789; the Corpus of Nineteenth-century English, which only covers the second half of the period and contains 1,800,000 words), all of which, as indicated above, are fairly restricted.

The CLMETEV is divided into three sub-periods (1710-1780, 1780-1850 and 1850-1920), each of which contains a different number of words. For this reason the figures obtained had to be normalized in order to achieve significant results. Also, whereas ahead is more frequent and it is present in a wider variety of texts, aboard, aloof and
*astern* are less frequent and more restricted in terms of the types of texts in which they appear. The total number of examples for the whole period is as follows: *aboard* 99, *ahead* 336, *aloof* 101 and *astern* 41. Table 1 shows the total number of instances and the normalized frequency (NF) per million corresponding to each period.

<table>
<thead>
<tr>
<th>PERIOD</th>
<th><em>Aboard</em> TOTAL</th>
<th>NF</th>
<th><em>Ahead</em> TOTAL</th>
<th>NF</th>
<th><em>Aloof</em> TOTAL</th>
<th>NF</th>
<th><em>Astern</em> TOTAL</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1710-1780</td>
<td>7</td>
<td>2.4</td>
<td>58</td>
<td>19.9</td>
<td>1</td>
<td>0.3</td>
<td>5</td>
<td>1.7</td>
</tr>
<tr>
<td>1780-1850</td>
<td>12</td>
<td>2.1</td>
<td>43</td>
<td>7.6</td>
<td>56</td>
<td>10.1</td>
<td>20</td>
<td>3.5</td>
</tr>
<tr>
<td>1850-1920</td>
<td>80</td>
<td>12.7</td>
<td>235</td>
<td>37.5</td>
<td>44</td>
<td>7</td>
<td>16</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Table 1. Number of examples found in the corpus.

All the documents in the *CLMETEV* are stored individually within each of the periods. This means that the collection process involved having to check all four words in each of the documents. In addition, these four words originally developed from the prepositional phrase “on + noun”, later “a + noun”, which implied that there might be three possible spelling representations for each of the words, namely, *aboard*, *a-board* and *aboard*. The three possibilities were checked in the corpus, which led to the gathering of the total number of examples mentioned above.

3. Results

3.1. *Meanings of the four terms found in the CLMETEV*

The four terms chosen for the present study show a wide range of meanings in the examples found in the *CLMETEV*, some of which are not included in the *OED*. Most of the original meanings of the words are also present in the examples extracted from the corpus, with the exception of *aloof*, which does not appear in any maritime contexts at all, as the figures show (see 3.2.). Although in most examples the four words had a clear meaning, some ambiguity was also found. Examples
were grouped according to their meanings and the results are shown below in the figures and tables in this section.

The meanings conveyed in the examples found in the CLMETEV are the following.\textsuperscript{11}

- **Aboard**
  a) Location on ships, boats or vessels (“he was aboard a ship”).
  b) Location on other means of transport (“all aboard for the Western cars”).
  c) Location inside the body* (“having more courage or more wine aboard than the rest”).\textsuperscript{12}

- **Ahead**
  a) Physical location in relation to the sea (“we saw the Land ahead”).
  b) Physical location in other places out of the sea (“four others, who had previously dashed on ahead on horseback”).
  c) Future time (“I pick up an independence for the days ahead”).
  d) Figurative location, that is, in relation to being advanced in abstract contexts (“she is ahead of us in knowledge”).

- **Aloof**
  a) Physical location in other places out of the sea, usually indicating where people are located in relation to other people (“the men all stood aloof to watch”).
  b) Figurative location, with the idea of being distant emotionally (“in order to be aloof from the discontent of the capital”).

\textsuperscript{11} The meanings are listed from the most original (the nautical meaning), physical and literal one to the most metaphorical one. Those that are not found in the OED are marked with an asterisk (*). The definitions belong to the author of the article and the examples accompanying each definition have been extracted from the CLMETEV.

\textsuperscript{12} This meaning is present in Wright’s dialectal dictionary (1961: 8), which clearly indicates this is a regional variety of a particular coastal area (Devon and Lincolnshire).
• *Atern*
  a) Location behind the ship, attached to the rear external part of the ship (“a spar which he towed astern”).
  b) Location about the stern, anywhere in that part of the ship (“to stand as far astern as possible and look over the side from the top deck”).
  c) Location behind the ship, not attached to it, at any distance as long as it is more or less on a direct line with the ship (“the iceberg had floated away astern”).
  d) Location behind other objects*. This particular meaning is only present in a few examples in relation to horses (“My companions, though seemingly gaily mounted, fell sadly astern”).

In addition to these meanings, some of the examples containing *ahead* and *aloof* could not be included in any of the groups because the meaning was considered ambiguous. When the full context was analysed, the meanings remained ambiguous and, therefore, they were included in a separate group, as shown in figures 2 and 3. One example of these is “I sent a Boat *a Head* to sound”, where “a boat” could have been sent physically in front or earlier in time, before the ones who where in the boat who sent it arrived at the place where it was sounding. Another example showing ambiguity is “he stood *aloof* from the Emperor and all his works”, where it is not clear whether “he stood” separated from the Emperor only physically and from what he was doing or whether “he” simply did not want to get involved with anything concerning the Emperor, which implies a more metaphorical meaning.

---

13 The *OED* does not provide any examples of *astern* being used in a non-maritime context and it does not clearly specify if *astern* could be used in other contexts. It simply says “in the rear (of a ship)”, which can be interpreted, from the use of parentheses, as if it could appear in other contexts, but is most likely to be used in relation to a ship.

14 The spelling shown in the examples extracted from the corpus may show some variation in relation to standard spelling. The reason for this is that the spelling has been kept as it appeared in the corpus.
Figures 1, 2, 3 and 4 show the presence of the different meanings in the three sub-periods of the corpus. Tables 2, 3, 4 and 5 include the total number of examples of each meaning in each period as well as the corresponding normalized figures (per one million words). As can be observed in general, whereas at the beginning of the period the literal original meanings of the words were more frequently present, towards the end of the period they had already achieved the range of meanings listed above.

<table>
<thead>
<tr>
<th>PERIOD</th>
<th>TOTAL (A)</th>
<th>NF (A)</th>
<th>TOTAL (B)</th>
<th>NF (B)</th>
<th>TOTAL (C)</th>
<th>NF (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1710-1780</td>
<td>7</td>
<td>2.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1780-1850</td>
<td>11</td>
<td>1.9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>1850-1920</td>
<td>66</td>
<td>10.5</td>
<td>14</td>
<td>2.2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Examples of aboard in the CLMETEV.

Figure 1. Distribution of the meanings of aboard found in the CLMETEV.

Figure 2. Distribution of the meanings of ahead found in the CLMETEV.

15 The bracketed letters correspond to the meanings that have been listed above. Tables 3 and 4 include a final group with the figures corresponding to the ambiguous examples (Amb).
Table 3. Examples of *ahead* in the CLMETEV.

<table>
<thead>
<tr>
<th>Period</th>
<th>TOTAL (a)</th>
<th>NF (a)</th>
<th>TOTAL (b)</th>
<th>NF (b)</th>
<th>TOTAL (c)</th>
<th>NF (c)</th>
<th>TOTAL (d)</th>
<th>NF (d)</th>
<th>TOTAL (Amb.)</th>
<th>NF (Amb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1710-1780</td>
<td>45</td>
<td>15.4</td>
<td>1</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>4.1</td>
</tr>
<tr>
<td>1780-1850</td>
<td>21</td>
<td>3.7</td>
<td>19</td>
<td>3.3</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>1850-1920</td>
<td>51</td>
<td>8.1</td>
<td>113</td>
<td>18</td>
<td>38</td>
<td>6</td>
<td>29</td>
<td>4.7</td>
<td>4</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Figure 3. Distribution of the meanings of *aloof* found in the CLMETEV.

<table>
<thead>
<tr>
<th>PERIOD</th>
<th>TOTAL (a)</th>
<th>NF (a)</th>
<th>TOTAL (b)</th>
<th>NF (b)</th>
<th>TOTAL (Amb.)</th>
<th>NF (Amb.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1710-1780</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1780-1850</td>
<td>28</td>
<td>5</td>
<td>22</td>
<td>3.9</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>1850-1920</td>
<td>19</td>
<td>3</td>
<td>14</td>
<td>2.2</td>
<td>11</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table 4. Examples of *aloof* in the CLMETEV.

Figure 4. Distribution of the meanings of *astern* found in the CLMETEV.
Table 5. Examples of *astern* in the *CLMETEV*.

As regards the types of texts in which they appear, their presence is restricted to very specific texts, mainly connected to the world of the sea in the early 18th century, whereas towards the end of the 19th century most of the texts included in the corpus contained at least one of the terms.

In addition, *aboard* and *ahead* also begin to appear in combination with other words, such as “all aboard” and “go ahead”. These collocations show a closer relationship at the end of the period, when they seem to be fully lexicalized as new lexical units and begin to acquire a more grammatical function, as for example the form “going ahead”, which appears functioning as an adjective (e.g., “It [the city of Buffalo, USA] is very going ahead”, which appears in *The Englishwoman in America*, written by Isabella Bird in 1856 and included in the *CLMETEV*).

3.2. Results word by word

3.2.1. Aboard

In the first sub-period the examples found in the corpus correspond almost exclusively to *Captain Cook’s Journal* (1768-1771). As the period advances, there is a clear growth in usage of all meanings. When other means of transport, such as the train, become common, during the 3rd sub-period (1850-1920), *aboard* began to be used for location outside a ship. Therefore, we can find examples such as “cars for Buffalo, all aboard”, which clearly refers to a different type of transport. This could be understood as a process of generalization,

16 All the examples included in this section have been extracted from the *CLMETEV*. 

---

<table>
<thead>
<tr>
<th>Period</th>
<th>TOTAL (a)</th>
<th>NF (a)</th>
<th>TOTAL (b)</th>
<th>NF (b)</th>
<th>TOTAL (c)</th>
<th>NF (c)</th>
<th>TOTAL (d)</th>
<th>NF (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1710-1780</td>
<td>2</td>
<td>0.6</td>
<td>1</td>
<td>0.3</td>
<td>2</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1780-1850</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>10</td>
<td>1.7</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>1850-1920</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0.6</td>
<td>11</td>
<td>1.7</td>
<td>1</td>
<td>0.1</td>
</tr>
</tbody>
</table>
since the usage of *aboard* is simply extended from one means of transport to another. The same would happen when other means of transport, such as the plane, become frequent and available to more and more people.

As regards the meaning found in relation to “in the body” (“having more courage or more wine *aboard* than the rest”), this is restricted to one single example and the extension seems to have been triggered by sociolinguistic aspects, more precisely, by the intake of alcoholic drink by sailors, particularly when they spent some time in the taverns in port areas. As mentioned above, the polysemous meaning of the noun *board* might have triggered these metaphorical extensions as well, since they may have developed from the meaning of *board* as the top of the table, where the drinks would have been located in those taverns, or from the fact that the sailors had bottles of alcoholic drinks on *board* ships, or even a combination of meanings. It is, however, essential to remember that this is not a very common expression and is restricted to some geographical dialects, as pointed out before.

Whereas the *OED* provides information about the usage of *aboard* in relation to other means of transport and the dates provided coincide in time with the examples found in the corpus (the first example found in the *OED* in relation to other means of transport dates from 1855 and the examples found in the corpus are all from 1850 onwards), there is no indication at all about the other metaphorical meaning found in relation to “having drink in the body”. This is probably due to its restriction in usage to specific dialects spoken in ports (Wright 1961: 8).

3.2.2. Ahead

Like *aboard*, in the first sub-period, *ahead* also appears almost exclusively in *Captain Cook’s Journal* (1768-1771). Similarly, it also experiences a clear growth in usage in all meanings in the third sub-period. From the second sub-period onwards an extension of the “physical location in relation to the ship” to other locations can be observed (e.g., “he [a mouse] gets so much *ahead* of them that he eats them up before they can again overtake him”). Again, this could be in-
terpreted as a case of generalization rather than as a metaphorical process as such.

As regards its temporal metaphorical extension, already in the 19th century there are some clear examples in the corpus that show that meaning (e.g., “I place myself at a distance of fifteen or twenty years ahead of this time”, which appeared in the novel *Confessions of an English opium-eater* (1822), written by Thomas de Quincey). The new meaning begins to appear in some ambiguous examples in the first sub-period. At the beginning of the LModE period there were some examples in the corpus which showed some ambiguity since they could either indicate ahead in space or ahead in time, such as in “send a boat ahead to sound” (Cook 1768-1771), as was described before. These ambiguous examples are essential to understand the evolution that the term experienced in relation to the development of a temporal meaning.

The first usage dates in the *OED* are different from our results, particularly as regards the temporal meaning, since the first example provided by the *OED* of ahead in relation to future time dates from 1900, whereas the first example with this sense found in the corpus dates from 1822, as mentioned above, leaving aside the ambiguous examples from the first sub-period, which were already pointing to that direction.

3.2.3. Aloof

The evolution of aloof is different from the evolution of the two terms examined so far. The first examples found in the corpus belong to the end of the 18th century. In addition to this, none of the examples was connected to the “sea”, that is, none of them referred to the original physical nautical meaning of the term. Basically, the meaning that is most frequently present in the corpus is ‘distant’ in relation to people, not only in a physical sense (e.g., “the ladies held aloof from her”), but also in a more metaphorical one (e.g., “Marcella held herself aloof and cold”). Once again, there seem to be a few ambiguous examples which could either mean ‘distant’ both in a physical and in a

17 These last two examples may look ambiguous to some extent, but the more extended context clearly indicates what they refer to.
metaphorical sense or simply one or the other, but even the more extended context does not clarify the definite meaning, as indicated above.

It is also a term whose usage decreased over the whole period of LModE. However, all the meanings that the OED provides seem to have been clearly represented in the examples found in the second sub-period (1780-1850).

The Industrial Revolution may have had an influence upon the evolution of this term. The arrival of the steam engine meant the decline in the importance of sails and the wind to move ships. As a consequence, terms that were highly connected to the wind and its influence on the sailing activity may have been affected by such an appearance, which is reflected in the decrease in usage of *aloof* along the LModE period.

The information found in the OED coincides with the information provided by the corpus in terms of the appearance of the new meanings, as well as the lack of examples particularly for the 17th and early 18th centuries. Since these terms had developed new meanings in a period earlier than the one represented in the corpus, a closer look at examples extracted from an EModE corpus would provide a clearer insight in the evolution of *aloof*.

### 3.2.4. Astern

The results of the examples found in the corpus containing the term *astern* show that it is a term whose usage decreases along the period. As opposed to the rest of the terms analysed in this study, it is the only one whose meanings are less metaphorical and the only one that is restricted to maritime contexts basically, as indicated above. It is also the only term whose meanings are all physical and refer to location or direction.

The fact that its original meaning gradually began to disappear in favour of a much wider meaning in relation to a position anywhere in the *stern* or behind the boat may have contributed to the appearance of the term in other contexts. However, the presence of *astern* in such metaphorical contexts is restricted to very few examples, all of them meaning “behind a horse”. As in the case of *aboard* used in relation to
other means of transport or ahead used to refer to location in front in other places out of the ship, this meaning of astern may also be understood as a process of generalization of the original meaning. The meaning of the noun, stern, has remained linked to the nautical context. Even though its meaning is very similar to ahead, though denoting the opposite direction of the ship, astern has not experienced the same kind of metaphorical extensions, related to time, as ahead has. One reason might be that the metaphorical meanings of the noun head were already fully established when it joined the prefix a-. Another reason is the fact that the word head is much more frequently used in a wider variety of contexts than stern, as shown by the whole range of meanings given for ahead in the OED. The first examples of head with a metaphorical meaning date from the end of the 14th century according to the OED, whereas the first example of ahead provided by the OED is from 1596. On the contrary, the first example of astern present in the OED dates from 1627, whereas the first and few metaphorical examples of stern given by the OED are from the end of the 16th century. Therefore, although this might imply that stern was already used metaphorically before it joined the prefix a-, the number of metaphorical examples is much smaller than that of the metaphorical examples of head, and in terms of time it does not take long for the prefix to join the noun. The writers of the novels from which the examples in relation to a horse were extracted may have simply used the term astern instead of “behind” for a matter of poetic licence and to call the reader’s attention.

The OED does not clarify whether the term astern is used exclusively in the nautical jargon or not. However, all the examples provided belong to a maritime context (e.g., “part of which they put into the boat, and the remainder they towed astern”). At the same time, the dates of all the examples given by the OED for the different meanings (17th and 18th century) are earlier than those of the examples found in the corpus, which explains the presence of most of the meanings already from the beginning of the LModE period.
4. Discussion

The evolution of the prepositional phrases from which the four chosen terms derived was very similar in the first stages, that is, the phrases became new lexical units due to a process of lexicalization. As has been explained, they developed different meanings and different syntactic categories, which appear to be the direct consequence of several processes of grammaticalization and metaphorical extension. However, they seem to have followed different patterns in this respect. Some of them, more specifically *ahead* and *aloof*, have acquired more abstract meanings, whereas *aboard* and *astern* seem to be much more restricted as regards the contexts in which they are used, while retaining a closer connection to their original physical meaning. Both the corpus and the *OED* confirm this. Also, the two sources point at the fact that the new meanings arose along the LModE period, with the exception of *aloof*, which had acquired a more abstract meaning in the previous period.

Nevertheless, there are some differences between the information contained in the *OED* and the data drawn from the examples found in the corpus, which are summarized in table 6.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>Aboard</th>
<th>Ahead</th>
<th>Aloof</th>
<th>Astern</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CLMETEV</strong></td>
<td>One example where <em>aboard</em> means ‘in the body’</td>
<td>First example referring to time dated 1822</td>
<td>No examples in a ‘physical maritime context’</td>
<td>Four examples ‘out of’ a maritime context</td>
</tr>
<tr>
<td><strong>OED</strong></td>
<td>No entry for ‘in the body’ meaning</td>
<td>First example referring to time dated 1900</td>
<td>All the meanings present in the examples of the corpus, as well as the ‘physical maritime’ one</td>
<td>No examples ‘out of’ maritime context and ambiguity regarding the existence of this meaning</td>
</tr>
</tbody>
</table>

Table 6. Differences between the *OED* and the CLMETEV.

The *OED* provides us with the first idea regarding all the meanings and the possible dates when the different meanings appeared. As the *OED* is a dictionary, it is not expected to provide ambiguous exam-
ples. However, as indicated before, in the case of *astern*, despite giving clear examples, the information in parenthesis might suggest ambiguity, i.e., there is a possibility that *astern* might appear in non-maritime contexts, but no examples of that are included. A lot of the new meanings, with the exception of *aloof*, seem to have appeared at the beginning of the LModE period if the attention is focused on the dates of the examples provided by the *OED*.

However, when observing the examples extracted from the *CLMETEV*, ambiguous examples often appear. This means that the whole context is required to clarify the definite meaning, although, unfortunately, the context does not always contribute to the complete clarification of the meaning of all the examples, which explains why some are still included in the ambiguous group. The examination of the examples confirms the appearance of the new meanings in LModE, and the fact that *aloof* had already become metaphorical in the previous period, since no examples of *aloof* with a physical, literal meaning in the nautical context were found.

5. Conclusions

Before stressing the importance of the use of corpora, it is necessary to refer to other findings that can be drawn from this study. It is important to take into account that some sociolinguistic factors may have influenced the evolution of these terms, namely the impact of the Industrial Revolution with the appearance of new means of transport and steam as a new power of propulsion, as well as the condemnation of the *a*-prefixed forms by 18th century grammarians. The Industrial Revolution may have contributed to the expansion of the meanings of *aboard*, since it began to be used in relation to other means of transport, such as the train, and also to the decline in usage of the term *aloof* with a physical nautical meaning, after the introduction of steam. As regards the influence of the 18th century prescriptive grammarians, once the condemnation clearly shown by them disappeared in the 19th
century, most *a*-prefixed terms began to appear much more regularly in all sorts of contexts, expanding their meanings and the variety of contexts in which they are present in the corpus, and showing greater usage, particularly from the second sub-period onwards.

Although there are not many unexpected results, since more abstract meanings are likely to appear later in time, some conclusions can be drawn from the analysis of the examples found in the corpus in comparison with the information offered by the *OED*.

It can be concluded that, whereas the *OED* does not provide ambiguous examples, they do exist in the corpus, and it is precisely these examples that provide very useful information for the analysis of the evolution of the different terms since ambiguity suggests the first stages of new meanings.

The corpus provided different results in relation to the dates of the appearance of some of the meanings. This indicates that these meanings were already present before the dates given by the *OED*.

Some of the meanings found in the corpus for these terms are not even mentioned in the *OED*, but their presence clearly means that they were used in the period of LModE.

Although the use of the *OED* offers a base for the analysis of meaning change, it does not provide a comprehensive evolution of meaning change, whereas the corpus clearly offers a closer insight into the reality of any given period, in this particular case the LModE period. Not only does it show a great variety of examples, but it also provides the whole context, which on some occasions proves essential to understand the real meaning of the example. Unfortunately, some ambiguous cases remain as such.

Because corpora give the context of word usage, as opposed to a dictionary’s precise definitions, these findings are not surprising. They confirm the importance of the use of corpora in any analysis of semantic change. Furthermore, they show the need for a greater variety of corpora of LModeE in order to carry out all sorts of linguistic analyses.

The 19th century examples extracted from the corpus also show a tendency towards the creation of lexical items in combination with some of these terms, namely, *all aboard* and *go(-)ahead*. Whereas the
former seems to have been fully established by that time, the latter is simply beginning to be established as one lexical unit.

To conclude, in addition to advocating the use of corpora in any kind of historical analysis involving changes both in meaning and grammar, this study shows that in the LModE period many language changes were influenced not only by linguistic inherent processes, but also by some sociolinguistic, cultural and historical developments of the period. It also confirms that words that originally appear in very specific contexts (the nautical jargon in this particular case) can develop new meanings that will probably be present in common usage before a dictionary records it. However, at present technological advances are opening new possibilities for newly coined words to appear in dictionaries online much sooner than they did when the second edition of the *OED* was published.

References


CLMETEV = De Smet (2006).


Dictionary-based corpus linguistics and beyond: developments in the expression of motion events in the history of English

Abstract

This chapter examines changes in the lexicon of English verbs of manner of motion (e.g., walk, crawl, dash, etc.) from Old English to the present day, with special attention to the emergence of new manner-of-motion construction types, such as the ‘sound emission to motion’ construction (SEtoM Cxn) (e.g., Sir Ascelin clanked into the hall). It also discusses some of the problems of data collection for the corpus linguist when faced with an open-ended linguistic system like vocabulary, as opposed to the more or less finite systems of grammar and phonology.

Author’s email address: teresa.fanego@usc.es; affiliation: U. of Santiago de Compostela. Some of the material in this paper was originally prepared for a plenary lecture presented at the 4th International Conference on Corpus Linguistics (CILC’2012), held at the University of Jaén in March 2012. I would like to thank participants at the conference for helpful discussions and the conference organizers for kindly inviting me to speak at such an intellectually rewarding event. For generous financial support, I am also grateful to the European Regional Development Fund and the following institutions: Autonomous Government of Galicia (Directorate General for Scientific and Technological Promotion, grant CN2012/012); Spanish Ministry of Economy and Competitiveness (grant FFI2011-26693-C02-01).
1. Introduction: sources of data

Over the past decades, the number of historical corpora available for the study of the English language has steadily grown. The compilation of the *Helsinki Corpus (HC)* (Kytö 1996 [1991]), the first comprehensive corpus to cover multiple genres from Old English (OE) to Early Modern English (EModE), was followed by the appearance of other widely used historical corpora, such as *A Representative Corpus of Historical English Registers (ARCHER, 1600-1999; see Biber et al. 1994; Yáñez-Bouza 2011)*, the *Innsbruck Computer Archive of Machine-Readable English Texts (ICAMET, 1100-1500; see Markus 2010)*, the *Corpora of Early English Correspondence (CEEC, 1418-1681)*, the *Corpus of Late Modern English Texts (Extended Version) (CLMETEV, 1710-1920)* and the *Corpus of Historical American English (COHA, 1810-2009)*, to name but a few.

In terms of size of corpora, there has also been a noticeable increase, from the 1.57 million words of the HC, the 1.78 million words of ARCHER and the 2.6 million words of CEEC to the nearly 15 million words of CLMETEV and the 400 million words of COHA. These two corpora in particular have largely served to solve the lack of scholarly attention that the English used in the Late Modern period (LModE; 1700-1900) had received until the turn of the millennium, at least in comparison with the language of earlier periods in the history of English.

Yet despite these evident developments in corpus compilation, for studies demanding a continuous historical record, or focusing on less frequent features and constructions and on lexical/semantic analysis, the researcher often has to make use of alternative – and

---

3 See <https://perswww.kuleuven.be/~u00444428>.
4 See <http://corpus.byu.edu/coha>.
5 This is the size of ARCHER in its original version, generally known as ARCHER 3.1. Since 2008 the corpus has been managed as an ongoing project by a consortium of participants at fourteen universities, among them Santiago de Compostela, and a new phase of expansion of the corpus is currently being carried out. For details, see <manchester.ac.uk/archer>. 
potentially less reliable – sources of data. Thus, for reasons that will become apparent later on, the research reported on in this article is not based on a corpus or corpora per se; instead I have opted for what one researcher famously called “dictionary-based corpus linguistics” (Mair 2004: 123). More specifically, I have resorted to two historical dictionaries, the Oxford English Dictionary (OED) and the Middle English Dictionary (MED) (Kurath et al. 1952-2001), with their large quotation databases covering more than a thousand years of English usage. These were supplemented with the Historical Thesaurus of the Oxford English Dictionary (HTOED) (Kay et al. 2009) and Chadwyck-Healey Literature Collections Online. Each of these resources will be briefly reviewed before addressing in §§2-4 the specific topic of this paper, namely developments in the expression of motion events in the history of English.

1.1. The Oxford English Dictionary (OED)

The OED is generally considered to be the world’s most comprehensive dictionary of the English language, currently covering the meaning, history, and pronunciation of 600,000 words – past and present – from across the English-speaking world. The first edition was published in fascicles between 1884 and 1928, first titled the New English Dictionary and later reprinted as the Oxford English Dictionary. This was followed by supplements in 1933 and 1972-1986 (in four volumes). A second edition, in twenty volumes, was completed in 1989, and was published in both print form (1989) and on CD-ROM (1992). A full-scale revision for a third edition (OED3) began in 1990, and revised entries have been published since 2000 and made available online (see Allan 2011: 18-20).6

As Hoffmann (2004: 18) notes, the compilers of the OED pursued the ambitious aim of illustrating the development of the form and meaning of each word with a series of quotations ranging from the first known occurrence of a word to the latest. In total, more than “five million quotations were collected for this purpose [...] and over 1.8

million of these quotations were used in the first edition of the *OED*. An additional 600,000 quotations were then added for the release of the second edition” (Hoffmann 2004: 18). Using the programme provided with the CD-ROM or with the online edition, this large collection of over 2.4 million quotations has become an important source of data in itself: it can be searched for individual lexical items or phrases and provides computerized lexical access to samples of the English language spanning more than 1,000 years and totalling about 25 million words (assuming an average length of ten words per citation – a very conservative estimate). The electronic format has thus revolutionized the way people use the dictionary to search and retrieve information. Complex investigations into word and phrase origins or quotations that would have been impossible to conduct using the print edition now take only a few seconds.

1.2. *The Middle English Dictionary (MED)*

The *MED* (Kurath et al. 1952-2001), completed in 2001 at the University of Michigan, has been described as “the greatest achievement in medieval scholarship in America”. Its 15,000 pages offer a comprehensive analysis of lexicon and usage for the period 1100-1500, based on the analysis of a collection of over three million citation slips, the largest collection of its kind available. The electronic version of the *MED* preserves all the details of the print *MED*, but goes far beyond this, by converting its contents into an enormous database, searchable in ways impossible within any print dictionary.

1.3. *The Historical Thesaurus of the OED (HTOED)*

The Historical Thesaurus of English project began in 1965 at the English Language Department of the University of Glasgow, with the aim of turning the headwords in the first edition of the *OED* into a conceptual thesaurus. Initially it was thought that the project would take

---

7 See <http://quod.lib.umich.edu/m/med>.
about fifteen years, but for a variety of reasons recounted in Kay (2011) it finally took 44 years for the work to reach the printed page (Kay et al. 2009).

In its final form, HTOED consists of the contents of the second edition of the OED, supplemented by OE vocabulary not included in the OED, all arranged in hierarchically structured conceptual fields containing lists of synonyms with their dates of use. Since 2010 HTOED has been linked to the online OED and thus revised in conjunction with OED3.

1.4. Chadwyck-Healey Literature Collections Online

Chadwyck-Healey Literature Collections Online is a virtual library containing over 350,000 literary texts, together with full-text journals, author biographies and other critical and reference resources. Since its launch in 1996, it has become established as a very popular tool among advanced researchers looking for accurate online versions of authoritative print texts.

For the linguist, the Literature Collections have the advantage of enabling quick searches for individual lexical items, as well as proximity searches for terms that have a specified distance between them. Thus, multiple occurrences of the so-called way-construction (e.g., *rustle all the way to the shore, crashing its way through the tree tops, rumbling its way upwards, splashed its way up the Hudson, etc.*) can be quickly retrieved by searching for the string *rustl* NEAR *way* and so on.

The individual Literature Collections that have been employed for this research are listed in the reference section.
2. Motion events across languages

The expression of motion events across languages has been a topic of lively debate since the publication, nearly forty years ago, of Talmy’s influential work (1975, 1985, 2000) on the classification of the world’s languages into satellite-framed and verb-framed. A motion event, according to Talmy (2000: II, 25-64), consists of four components: a) a figure moving with respect to another entity; b) the reference entity, or ground, with respect to which the figure moves; c) the path followed by the figure with respect to the ground; d) the motion. Thus in (1) below, he functions as the figure, the room as the ground, into expresses path, and went motion.

(1)  He went into the room.
Figure  Motion   Path  Ground

Motion events can be simple or ‘unitary’ as in (1), which indicates only one dimension of the motion (in this case, the path information, i.e., into), and complex as in (2-4). A complex motion event encodes or ‘conflates’ within a single clause an additional component of motion, very often the manner. Thus a manner, in addition to a motion, is expressed by walked in (2), floated in (3), and crawled in (4):

(2)  She walked in.
(3)  The bottle floated out (of the cave).
(4)  She crawled there.

Talmy’s concern is with how the various components of a motion event are realized in surface expression. He pays attention in particular to two classes of surface elements, verbs and satellites, and their patterns of lexicalization or ‘regular association’ with one or more of those meaning components (cf. Talmy 1985: 59). Satellites (Talmy 2000: II, 102) can be either bound affixes or free words, and encompass a number of grammatical forms which have traditionally been treated independently of each other, such as English verb particles, German separable and inseparable verb prefixes, and Latin and Russian verb prefixes. In this paper, the inventory of English satellites
will be considered to include verb particles proper (in, out, off, along, around, etc.), prepositional and adverbal phrases (PPs, AdvPs) (e.g., into the room in (1), there in (4)), and elements such as together, apart and forth.\(^8\)

Based on how the component path is mapped onto syntactic structure in a complex motion event, Talmy (1975, 1985, 2000) claimed that the world’s languages generally seem to divide into two distinct classes: satellite-framed languages (S-languages) such as English, German, Swedish, Russian and Chinese, which characteristically encode manner in the main verb and path in a satellite (as in examples (2)-(4) above), and verb-framed languages (V-languages) such as Spanish, French, Turkish and Semitic, which characteristically express path in the verb, and leave the expression of manner to an independent, usually adverbal or gerundive type constituent whose presence in the clause is optional and, indeed, is often not expressed at all. Witness the following examples from Spanish:

(5) La botella salió de la cueva (flotando).
the bottle moved-out from the cave (floating).
‘The bottle floated out of the cave’.

(6) Entró en el cuarto, acompañada del susurro siseante de sus ropas.
she entered the room, accompanied by the rustle of her clothing.
‘She rustled into the room’.

Subsequent work on motion events has led to a revision of Talmy’s typology in a number of ways (cf. Slobin 2004a: 249ff; Beavers et al. 2010; Croft et al. 2010; Croft 2012: 292-319), and also to a growing interest in the diachrony of motion event systems (cf. Kopecka 2006; Masini 2006; Iacobini/Masini 2007; Iacobini 2012; Verkerk forthcoming, etc.) and the effects of the linguistic differences between S-framed languages and V-framed languages on the shape and development of verb lexicons. This latter aspect in particular has been investigated by Slobin and colleagues (among others, Berman/Slobin 1994; 2004a; 2006).

---
8 PPs and AdvPs are not considered satellites in Talmy’s categorization, but, as has often been pointed out (cf. Narasimhan 2003: 149-150; Beavers et al. 2010: 339; Croft et al. 2010: 205-206), they serve the same function as satellites proper in indicating path-related notions.
Slobin 1997, 2004a, 2004b, 2006a: 71-72, 2006b). In a series of seminal papers, Slobin has compared translations of narratives in different languages as well as the linguistic behaviour of speakers of different languages in a variety of tasks involving the description of pictorially presented motion scenes. These analyses have shown, *inter alia*, that speakers of S-languages tend to describe motion with manner verbs, while speakers of V-languages predominantly use path verbs. In addition, or perhaps as a result of the previous observation, speakers of S-languages mention manner of motion more frequently and with greater lexical diversity. Building on these results, Slobin has put forward the idea that such linguistic differences are in turn likely to have effects both on the organization of mental representations and on the lexicon of the language in question, and has proposed (2004a) a *diachronic model for the emergence of manner salience*. As he puts it, “if manner is easily accessible it will be encoded more frequently and, over time, speakers will tend to elaborate the domain in terms of semantic specificity. Consequently, learners will construct a more elaborate conceptual space for manner, allowing each new generation to continue the cycle of attention to manner” (Slobin 2004a: 246). Hence, Slobin (2006b): “the semantic fields of frequently accessed lexical items become salient conceptual domains. Such domains attract new lexical items, filling slots in entrenched construction types”.

In an earlier paper (Fanego 2012), I tested Slobin’s hypothesis on the emergence of manner salience by examining developments in the encoding of motion events from OE to LModE (1700-1900). My findings showed that, in accordance with Slobin’s predictions, the English domain of manner of motion has constantly been on the increase, with impressively large additions of innovated and borrowed forms in each of the historical periods (OE, ME, EModE and LModE). These findings are briefly summarized in §3; after this I will consider in §4 the related question, not previously discussed, of the extent to which the continued attention to manner may have led not just to the saturation of the domain over time, but also, crucially, to the emergence of new manner-of-motion construction types.
3. The conceptual domain of ‘manner of motion’ from OE to LModE: elaboration in terms of semantic specificity

As already noted, English, a prototypical S-language, has a large number of verbs lexicalizing motion in general, and manner of motion in particular. A definitive count of manner-of-motion verbs in English has yet to be undertaken, but Slobin (2004a, 2006a: 71, and elsewhere) gives comparative figures for several languages, which his team obtained by means of back translations, dictionary searches and corpora: “the satellite-framed languages [...] English, German, Dutch, Russian and Hungarian each have several hundred manner verbs; Mandarin has perhaps 150; Spanish, French, Turkish, and Hebrew have less than 100, and probably less than 60 in everyday use”.

In view of this abundance of manner-of-motion verbs in contemporary English, in Fanego (2012) I sought to make the data manageable by restricting the analysis to self-agentive manner verbs of translational motion (cf. Talmy 2000: II, 25, 28), that is, intransitive manner verbs describing a change in the location of the figure and having animate agents as their grammatical subjects, an example being *I ran down the stairs*. This implies that the following classes of motion verbs were excluded:

a) Transitive verbs of caused manner of motion: *I slid the keg into the storeroom*.

b) Non-agentive verbs of self-contained motion (motion consisting of oscillation, rotation, dilation, and the like), such as *float, roll* or *bounce*, with which “there is not necessarily protagonist control on the part of the moving entity” (Levin 1993: 265).

c) ‘Neutral’ motion verbs (cf. Özçalişkan 2005: 213), which encode neither path nor manner: *go, move*.

d) Path verbs, which include a specification of the direction of motion, but not manner (also often referred to as verbs of inher-
ently directed motion; cf. Levin 1993: 263-264): enter, exit, ascend, descend, leave, etc.\(^9\)

With regard to the procedure for data collection, the open-endedness of the lexicon, as opposed to the more or less finite systems in grammar or phonology, poses a problem for the identification of the resources of the English language for describing manner of motion. Technically, the ideal strategy would have been to proceed entry by entry through historical dictionaries and dictionaries of contemporary English. But as this was clearly not feasible, the heuristic strategy I followed consisted of the following two steps:


b) I then looked up these words in dictionaries and thesauruses, so as to find further candidates for inclusion. For earlier stages of English I conducted searches using the sources discussed in §1 above; they were supplemented with searches in other historical dictionaries (B&T, T&C, DOE), the Thesaurus of Old English Online (Edmonds et al. 2005), and a number of earlier publications on the topic (e.g., Ogura 2002).

In this way, I was able to obtain the results summarized in Table 1. They confirm a continued attention to the manner-of-motion lexicon by English language users, with dozens of new manner verbs added over the course of time, many of them belonging to what Slobin (1997: 459) has termed ‘second tier manner verbs’, that is, expressive or exceptional verbs denoting very specific kinds of motion, such as

---

\(^9\) Manner verbs may also encode path, and in such cases were counted as regular manner verbs, e.g., sky-rocket ‘(of persons) to fly or jump up suddenly, in the manner of a sky rocket’ (OED s.v sky-rocket v. 2.b). The same applies to verbs such as flee, plunge and tumble, which Levin (1993: 263) includes in her class of inherently directed motion.
**Dictionary-based corpus linguistics and beyond: motion events in English**

...dash, swoop, scramble (see also the lists in (7)-(9) below), and thus making distinctions that do not play a role in the considerably smaller manner lexicons of V-languages. These developments seem fully in accordance with Slobin’s predictions on the emergence of manner salience; as he observes, “as a domain becomes more saturated in a language, speakers invent lexical items to mark distinctions that become important to them. A language with a rich manner lexicon tends to get richer over time [...] leading to further lexical innovation” (Slobin 2006a: 71-72).

<table>
<thead>
<tr>
<th>TIME PERIOD</th>
<th>NUMBER OF INNOVATED MANNER VERBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 1100</td>
<td>72</td>
</tr>
<tr>
<td>1100-1500</td>
<td>181</td>
</tr>
<tr>
<td>1500-1700</td>
<td>205</td>
</tr>
<tr>
<td>1700-1900</td>
<td>250</td>
</tr>
</tbody>
</table>

Table 1. English intransitive verbs of translational manner of motion. Number of verbs added per period.

(7) Examples of innovated manner-of-motion verbs in ME:

- native verbs: *burst* ‘to rush violently’; *clamber* ‘to climb with difficulty’; *crowd* ‘to press on, hurry’; *foot* ‘to walk or go on foot’; *roam* ‘to wander’; *steal* ‘to come or go secretly’; *wade* ‘to walk through water or any liquid’; *walk* ‘to go on foot, walk’; etc.
- loans from Old French: *coast* ‘to sail along the coast; to travel across’; *dance*, *haste*, *jet* ‘to strut, swagger’; *journey*, *launch* ‘to leap, rush, charge’; *march* ‘to march against somebody, attack’; *renge* ‘to roam’; *rush*, *trail* ‘to walk with long trailing garments’; *travel*, etc.
- loans from Old Norse: *crawl* ‘to walk slowly’; *dash*, *fling* ‘to rush, dash’; *lurk* ‘to go stealthily’; *skulk* ‘to go or move furtively, to slink’; *blunder* ‘to walk blindly’; *thrust* ‘to push one’s way, jostle’; etc.
- loans from Middle Dutch or Middle Low German:¹⁰  *hobble* ‘to walk lamely, hobble’; *tramp* ‘to tread heavily, stamp’; *whip* ‘to move back and forth quickly’; etc.

---

¹⁰ Apart from French, Scandinavian and Latin the only other substantial foreign sources to influence ME lexis directly were the languages of the Low Countries (Flemish, Dutch and Low German), which, like English, are strongly S-framed languages. Borrowings, partly through commercial and military contacts, and partly by the settlement of Flemish weavers and farmers in England
Examples of innovated manner-of-motion verbs in EModE:

- semantic extensions of already existing verbs: *daggle* ‘to walk in a slovenly way (through mud or mire), to drag or trail about’; *file* ‘to march or move in a file’; *jog* ‘to walk or ride with a jolting pace’; *mince* ‘to walk with short steps in an affected manner’; *pad* ‘to journey on foot, esp. as an itinerant’; *perambulate* ‘to wander or travel from place to place’; *peruse* ‘to continue (a journey), to travel, to wander’; *procession* ‘to go in procession’; *rustle* ‘to move with a rustling sound’, etc.

- new formations (by conversion or other word formation methods): *barge* ‘to journey by barge’; *career* ‘to take a short gallop, pass a “career”’; *chariot* ‘to ride or drive in a chariot’; *clutter* ‘to crowd together’; *gallop* ‘to ride at full speed’; *hoof* ‘to go on foot’; *inch* ‘to move by inches or small degrees’; *lavolta* ‘to dance a lavolta’; *lighter* ‘to remove or transport (goods) in a lighter, or as in a lighter’; *progress* ‘to make a progress or journey’; *promenade* ‘to take a leisurely walk’; *scull* ‘to proceed by means of a boat propelled with a scull’; *skate* ‘to glide over ice upon skates’; *wayfare* ‘to journey or travel, esp. on foot’, etc.

Examples of innovated manner-of-motion verbs in LModE:

- semantic extensions of already existing verbs: *blow* ‘to move as if carried or impelled by the wind’; *circulate* ‘to go about in a social circle, spec. to move around at a gathering’; *clank* ‘to move with a clanking sound’; *overstride* ‘to take longer strides than is natural’, etc.

- new formations (by conversion or other word formation methods): *cancan* ‘to dance the cancan’; *canoe* ‘to paddle or propel a canoe’; *cavalcade* ‘to ride in a cavalcade, esp. in procession or in company with others’; *constitutionalize* ‘to take a “constitutional” (walk)”’; *ski* ‘to travel on skis’; *sledge* ‘to travel in a sledge’; *sleep-walk*; *toboggan* ‘to ride on a toboggan or sleigh’, etc.

and Wales, began quite early, extending “from the days of William the Conqueror, whose wife was Flemish, down to the eighteenth century” (Baugh/Cable 2002 [1951]: 187).
4. The emergence of new manner-of-motion construction types

4.1. Theoretical preliminaries

Cognitive Linguistics (Lakoff/Johnson 1980; Lakoff 1987; Langacker 1987, 1991; among many others) proposes a usage-based model for language use, language acquisition and language change. This usage-based model contrasts with the archetypal conception of language as a system of general rules characteristic of generative models, where “irregular and idiosyncratic phenomena cannot therefore be accommodated in a natural or convincing manner” (Langacker 1987: 46) and are either ignored in the quest for generalizations or listed in the lexicon. In a usage-based model, on the other hand,

particular statements (specific forms) coexist with general statements (rules accounting for those forms) in a speaker’s representation of linguistic convention, which incorporates a huge inventory of specific forms learned as units (conventional expressions). Out of this sea of particularity speakers extract whatever generalizations they can. Most of these are of limited scope, and some forms cannot be assimilated to any general patterns at all. Fully general rules are not the expected case in this perspective, but rather a special, limiting case along a continuum that also embraces totally idiosyncratic forms and patterns of all intermediate degrees of generality. The archetypal conception is thus seen as a matter of false expectations. (Langacker 1987: 46)

These developments in cognitive and usage-based linguistics have afforded new perspectives on language change at all levels. As Bybee (2007: 946) notes,

[the rejection of the rule/list fallacy in favor of usage-based exemplar storage as proposed in the work of Langacker provides a grammar that is more compatible with the lexical and phonetic gradualness of change, including not just sound change, but also analogical change, grammaticization, and syntactic change.

Referring specifically to the paths of syntactic change, many historical linguists, in particular those writing from the perspective of grammaticalization, have observed that the emergence and growth of a construction proceeds in an incremental fashion, with changes being local
and gradual: “[g]rammar is built up from specific instances of use that marry lexical items with constructions; it is routinized and entrenched by repetition and schematized by the categorization of exemplars” (Bybee 2006: 730; emphasis added).

In accordance with this usage-based model of language use and linguistic structure, the model of grammatical analysis known as Construction Grammar\(^{11}\) proposes to expand the role of the traditional lexicon to include productive or semi-productive phrasal patterns (the ‘constructions’) that had previously been assumed to lie within the domain of syntax. Essentially, constructions are stored pairings of form and function “connected to meaning in a conventionalized and partially idiosyncratic way” (Goldberg/Jackendoff 2004: 532-533; also Goldberg 1994: 4, and elsewhere); in addition, regularities between form and meaning are also considered constructions, even if they are fully compositional, “as long as they occur with sufficient frequency” (Goldberg 2006: 5).

Also central to Construction Grammar, and more specifically to Lakoff and Goldberg’s version of it, are notions such as motivation and inheritance. In Lakoff’s early formulation (Lakoff 1987: 65, 91, 537 ff; also Goldberg 1995: 69 ff; Goldberg/van der Auwera 2012: 110-111), a given construction is motivated by its mother construction(s) to the extent that shared form and function are inherited from the mother construction(s). Motivational relationships can be made explicit “via a default inheritance hierarchy, which represents how our knowledge of language is not a set of unrelated idioms, but rather a network of related constructions” (Goldberg/van der Auwera 2012: 129). Incorporating motivation into the grammar thus captures the fundamental structuralist insight, overlooked by most formal linguistic theories, that “elements in a system influence each other even when they do not literally interact [...] speakers (unconsciously) seek out

---

\(^{11}\) The label Construction Grammar (capitalized) is used here primarily for the version of construction grammar developed by Goldberg (1995, 2006). Other ‘constructional’ theories of grammar include the Berkeley Construction Grammar of Fillmore, Kay and collaborators (Fillmore/Kay 1993; Kay/Fillmore 1999, etc.), Cognitive Grammar (Langacker 1987, 1991) and Radical Construction Grammar (Croft 2001).
regularities and patterns, and tend to impose regularities and patterns when these are not readily available” (Goldberg 1995: 72).

In what follows, I will briefly illustrate these various aspects by means of a case study, namely the pattern that I will refer to henceforth as the ‘sound emission to motion’ construction (SEtoM Cxn). Its emergence in the history of English is to be understood against the background of the diachronic model for the emergence of manner salience proposed by Slobin (2004a) discussed earlier in this paper.

4.2. The ‘sound emission to motion’ construction (SEtoM Cxn)

English has a class of verbs which, following Levin/Rappaport Hovav (1991: 138), I will label verbs of sound emission. A list of approximately 120 verbs in the class is given in Levin (1993: 235); a few illustrative examples follow here:

(10) babble, bang, beep, buzz, chatter, clang, clash, clink, groan, growl, jingle, murmur, rattle, ring, splash, squeal, sing, thunder, trumpet, ululate, whine, whistle

Semantically, “these verbs all describe an event in which something or someone emits a sound: the core sense of the verb” (Levin et al. 1997: 27), as in (11)-(12):

(11) The machine rumbled as it started up. (OED rumble v.2 3. ‘To make a low heavy continuous, but varying, sound’)
(12) How that door creaks! (OED creak 2.a.intr. ‘To make a harsh shrill grating sound, as a hinge or axle turning with undue friction’)

It is a characteristic of some verbs of sound emission, in English and other languages (see e.g., Levin/Rappaport Hovav 1995: 202 ff on German and Modern Hebrew; Verkerk forthcoming: Appendix 3a, on Dutch and Swedish), that they can be used in various constructions encoding motion. These include (labels for each pattern as in Levin et al. 1997) the causative motion construction (the wind creaked the door to and fro); the way-construction (the coach rumbled its way for miles);
the resultative construction (*the dog barked the intruder back to the front door*); and the intransitive sound emission to motion construction\(^{12}\) (*the cab rumbled back to town*). Sound emission verbs are not, however, uniformly distributed over these four patterns; in other words, according to Levin et al. (1997: 39-40), only the *way*-construction seems to be fully productive and available with all sound emission verbs.

From the diachronic point of view, the *way*-construction is also the only one to have been investigated in some detail, initially by Israel (1996) and more recently by Mondorf (2011) and Traugott/Trousdale (2013). Its emergence in the history of English, variously dated back to the 15th century by Israel (1996) and to EModE by Traugott/Trousdale (2013), seems to have taken place at about the same time as the emergence of the construction which will be my concern in the remainder of this paper, namely the intransitive SEtoM Cxn, as in *the cab rumbled back to town*. A couple of 19th-century examples of this are given here:

(13) 1868 J. G. Whittier *Among Hills* Prel. 13: A single hay-cart down the dusty road Creaks slowly. (*OED* creak v. 2.c ‘To move with creaking’)

(14) 1866 C. Kingsley *Hereward* xxi. 270: Sir Ascelin clanked into the hall. (*OED* clank v. 3 ‘To move with a clanking sound’)

4.2.1. Formal and semantic properties of the SEtoM Cxn

As in Cognitive Linguistics generally, I assume that constructions are ‘schematic’ (i.e., abstract and not specified in detail) pairings of form and meaning (Langacker 1987: 58, and elsewhere). More specifically, for the intransitive motion construction (*He walked into the room*) discussed in §3 above and exemplified in (1)-(4) I will propose, following Goldberg/Jackendoff (2004: 540), the semantic and syntactic structure in (15); the subscripts indicate the correspondence between the syntactic arguments and the corresponding semantic arguments; the term ‘co-event’ is used to refer to the additional component of

---

12 This is my own label, replacing Levin et al.’s (1997) label ‘intransitive motion accompanied by sound’.
motion which is conflated in complex motion events (on the structure of these see §2 above):

(15)  Intransitive motion construction \((He \text{ walked into the room}; \text{ The ball rolled down the hill})\)
Syntax:  \(\text{NP}_1 \text{ V } \text{PP}_2\)
Semantics:  \(X_1 \text{ MOVE } \text{Path}_2\)
co-event:  MANNER
i.e., ‘walking is the manner of moving into the room’; ‘rolling is the manner of moving down the hill’.

This is a pattern of co-event conflation (Talmy 2000: II, 27) common in many languages: the verb expresses at once both the fact of motion and a co-event of manner describing the means of propulsion that brings about the motion. In English, this type of construction is already attested in OE, and has since been employed both with increasing token frequency and with an increasing number of verbs (type frequency); see Fanego (2012) and Table 1. The following are early examples:

(16)  \textit{Aelfric’s Lives of Saints} (Swithun) 100: \textit{he aras \text{\`a} on mergen…, and mid twam criccum creap him to Wynceastrre ‘he rose then in the morning…, and with two crutches crawled to Winchester’ (DOE creopan 1b.i)}
(17)  \textit{Homilies of Aelfric} 5 214: \textit{þæt wif … efste to ðære byrig and bodade ymbe Crist ‘the woman … hurried to the city and preached about Christ’ (DOE ef(e)stan A.1.a)}
(18)  \textit{c1450 Alph.Tales} (Add 25719) 183/12: \textit{He went furth vnto þe myln-dam of þe abbay, & þer he lowpid in & dronymd hym. ‘He went forth to the mill-dam of the abbey, and jumped therein and drowned himself’. (MED lopen (a))}

For the intransitive SEtoM Cxn which is my concern here, I will assume, by contrast, the slightly different structure in (19); this captures the fact that, as Croft notes (2012: 302), “it is the motion (including its manner) that brings about the sound emission”: 
Intransitive ‘sound emission to motion’ construction (*The cab rumbled back to town*)

Syntax: \[ NP_1 \ V \ PP_2 \]

Semantics: \[ X_1 \ MOVE \ Path_2 \]

co-event: CONCURRENT RESULT

i.e., ‘the rumbling is a result of the cab’s motion back to town’

I would argue that the emergence of this structure in English is ‘motivated’, in the sense discussed in §4.1 above (cf. also Lakoff 1987: 65, 91, 537 ff; Goldberg 1995: 69 ff; Goldberg/van der Auwera 2012: 110-111), by the existence of the closely related structure represented in (15) and illustrated in (16)-(18).

### 4.2.2. Growth and history of the SEtoM Cxn

In order to examine the history of the pattern in (19) I began with Levin (1993: 235), who provides a list of approximately 120 verbs of sound emission. Not all of them, however, can be used in a motion sense (among others, *crackle* ‘to emit a rapid succession of slight cracks’, *squawk* ‘to call or cry with a loud harsh note’, etc.), so my first step was to check how many are actually available today in the pattern under discussion. This far shorter list of verbs thus obtained was then used to look for synonyms and further candidates for inclusion in the *MED*, *OED* and *HTOED*. In this way I ended up with a final list of some 77 sound verbs capable of occurring in motion constructions; many of them are still in use in PDE, but a few have become obsolete.

Overall, the intransitive SEtoM Cxn appears to develop in accordance with the model of language change outlined in §4.1, with the new construction emerging from a few often highly specific instances and then expanding in its own direction.  

---

13 In this regard note also Croft/Cruse (2004: 326): “specific instances of construction schemas [...] function as ‘islands’ from which a new construction expands, establishing and generalizing a new construction schema with its own syntactic and semantic peculiarities”.
The earliest apparent instances I have been able to find so far occur with frequentative verbs in -er (Germanic -rôjan) and -le (Germanic -ilôjan), two suffixes forming disyllabic verbs expressive of (repeated) sound or movement (cf. Marchand 1969 [1960]: 273, 322-323; also OED s.v. -er suffix5, -le suffix 3), and thus it seems that phonetic motivation has played a part. The specific verbs involved at this stage are only clatter, rustle and rumble:

(20) a1450(a1400) Siege Jerus.(1) (LdMisc 656) 569: Castels clateren doun. (MED clateren v. 1.(b). ‘To break or fall with loud noise’)
(21) (a1470) Malory Wks. (Win-C) 736/3: He smote hym so harde uppon the shylde that sir Palomydes and his whyght horse rostele to the erthe. ‘He struck him so hard upon the shield that Sir Palomides and his white horse rustled to earth’. (MED rustelen v. (b). ‘To fall noisily’)
(22) c1450(c1380) Chaucer HF (Benson-Robinson) 1026: .. heren..The grete soun..that rumbleth up and doun In Fames Hous, full of tydynges, ‘hear... the loud noise... that rumbles up and down in the House of Fame, full of tidings’, (MED rumbelen v. (d). ‘To move with a rumbling sound’)

In terms of syntactic structure, all three examples are indistinguishable from both the intransitive motion construction exemplified in (15)-(18) above and from the SEtoM Cxn proper. Semantically, however, there exist important differences. Thus, (20) and (21) do not encode translational motion (Talmy 2000: II, 25, and elsewhere), that is, motion “in which the location of the Figure changes in the time period under consideration” and traverses the path expressed by the prepositional phrase. Witness the impossibility, for instance, of glossing (20) as ‘the clattering is a result of the castles’ motion (along the specified path). Rather, they seem to be instances of Talmy’s self-contained motion, in which “an object keeps its same basic, or ‘average’, location. Self-contained motion generally consists of oscillation, rotation, dilation (expansion or contraction), wiggle, local wander, or rest” (Talmy 2000: II, 35-36).

As regards (22), a paraphrase such as ‘the rumbling is a result of the sound’s motion up and down’ is equally inappropriate. Indeed, it could be argued that this is not a motion sentence at all, as the ‘sound’ does not really traverse the path, that is, there is no change in the location of the figure (just as the road runs along the coast, for
instance, does not qualify as a motion sentence, except meta-
phorically).

Uses of this kind, however, were probably the starting point (or one of the starting points) for unambiguous instances of the SEtoM Cxn. In my data, these are found from EModE times, and with only a handful of verbs, including those mentioned earlier, and a few others of related meaning, such as buzz and thunder:

(23) 1594 J. Lyly Mother Bombie v. iii. sig. H3: These minstrelles... rustle into
every place. (OED rustle v. 2.a. ‘To move with a rustling sound’)

(24) 1604 T. M. Blacke Bk. in T. Middleton Wks. (Bullen) VIII. 28: I thumped
down stairs with my cowheel. (OED thump v. 3.b. ‘To walk with heavy
sounding steps, to stump noisily’)

(25) 1642 J. Taylor St. Hillaries Teares 5: The Coaches which had wont to rumble
up and downe. (OED rumble v. 2 4.a. ‘To move or travel with a rumbling
sound’)

(26) 1665 Richard Head The English Rogue, Part 1, 152: and if alone, he would so
thunder down the stairs, (fear giving wings to his feet) as if (Vulcan like) he
had been sent by Jupiter head-long in a message. (Chadwyck-Healey, Early
English Prose Fiction). Cf. OED thunder v. 2.a. ‘To rush or fall with great
noise and commotion’

(27) 1697 Dryden tr. Virgil Æneis xi, in tr. Virgil Wks. 573: When the Jav’lin
whizz’d along the Skies. (OED whizz v. 2.a. ‘To move swiftly with or as with
a sound as of a body rushing through the air’)

Over the course of the 18th century the construction expanded to some 10 additional verbs (see 28-29), and in the 19th to a further, much larger group of 31 verbs. My material for these two centuries reflects well the multiple motivation that often shapes actual processes of change over an extended period of time: in principle, the SEtoM Cxn is analogically extended to verbs which denote sound and are thus semantically near enough to other verbs already occurring in the construction. But in addition, the process also seems to have been driven by factors such as phonetic symbolism (see, among others, Marchand 1969 [1960]: Chapters VII and VIII; Samuels 1972: 45 ff; Hinton et al. 1994). Thus plash is first found in the sense ‘to move through water splashily’ in 1577 (in Chadwyck-Healey, Early English Prose Fiction, Grange, The Golden Aphroditis 1577: 73), and throughout the 17th and 18th centuries seems to have been common in this use. This
may have facilitated the expansion of the SEtoM Cxn, during the 19th century, to a number of verbs of highly specific meaning also ending in the ‘phonaesteme’ /-ʃ/ (Marchand 1969 [1960]: 400): clash ‘to move with violence and noise’, crash ‘to move or go with crashing’, hish ‘to move with a rushing or whishing noise’, slosh ‘to move splashing about in mud or wet’, slush ‘to rush (down) with a splashing sound’, splash ‘to move with a splash’, squash ‘to move or walk with a splashing sound’, swish ‘to move with the sound expressed by ‘swish’’, swoosh ‘to move with the a noise expressed by the syllable ‘swoosh’’, whish ‘to move with a soft sibilant sound resembling the syllable ‘whish’’, etc.

(28) 1748 Samuel Richardson, Clarissa (1st edition), p. 337: These the noises, and the speeches, as we clatter’d by the door of the fair briberess. (Chadwyck-Healey, Eighteenth-Century Fiction)
(29) 1751 Moses Mendez, The shepherds lottery, p. 23: While adown the slopy Hill /Tinkles soft the gushing Rill, (Chadwyck-Healey, English Drama)
(30) 1847 Ld. Haddo Mem. (1866) I. 16: We... slosh through the moor to a shepherd’s house. (OED slosh v.1 1. ‘To splash about in mud or wet’)
(31) 1826 Timothy Flint, Francis Berrian; or, The Mexican Patriot, Vol. 1, p. 142: We now passed through deep and still fore sts; then splashed through swamps and streams. (Chadwyck-Healey, Early American Fiction)
(32) 1859 Dickens Tale of Two Cities i. ii. 4: Once more, the Dover mail struggled on, with the Jack-boots of its passengers squashing along by its side. (OED squash v.1 3. ‘To move or walk with a splashing sound’)
(33) 1820 Charles Robert Maturin, Melmoth The Wanderer, p. 285: I would as soon be on the Indian seas, where your dreams send me rowing every night, or crashing through the ice near the Poles... (Chadwyck-Healey, Nineteenth-Century Fiction)

With regard to the 20th century, although this is a period I have not yet investigated in detail, my data so far suggests that the SEtoM Cxn has continued to expand: between 1900 and 1990 around 23 new verbs are recorded in the pattern, according to my sources for this period (essentially the OED3 and HTOED). The SEtoM Cxn now has a high enough type frequency and a sufficiently coherent semantics that it ‘sanctions’ (Langacker 1987: 66) the creation of novel instances no longer necessarily dependent on the prior existence of a sound verb of the appropriate kind; for, as Croft notes (2012: 303), recurrent aspects
of meaning sanction the use of a construction “in new situations that the speaker wants to describe”. It is worth noting, for instance, that many 20th-century instances of the SEtoM Cxn appear simply to serve as a way of offering verbhood to nouns, interjections and other lexical items imitative of the sound produced by machines and engines of various kinds:

(34) 1929 J. B. Priestley Good Compan. 612: The train... slowly **chuff-chuffed** into the gloom. (*OED* chuff v.² intr. Of an engine or machine: ‘to work with a regularly repeated sharp puffing sound’. Also n.)

(35) 1939 I. T. Sanderson *Caribbean Treasure* ix. 176: During the half-hour that we had been **‘phut-phutting’** up the Surinam to the polder, both Fred and I had learnt many things about our craft. (*OED* phut-phut v. [< phut int. and n.] ‘To make intermittent muffled explosive sounds like a small internal-combustion engine; to move with such a noise’)

(36) 1917 ‘CONTACT’ *Airman’s Outings* 114: Slowly... our train **chugged** northward. (*OED* chug v. ‘To move with a sound characteristic of a steam-engine or electric motor at work’)

(37) 1975 D. Lodge *Changing Places* 190: Morris pulled out and **varoomed** down the wrong side of the road. (*OED* vroom n. ‘The roaring noise of a motor vehicle accelerating or travelling at speed. Also as v. intr., to make such a noise; to travel or accelerate at speed’)

5. Conclusions

The clearest finding from the data adduced here is that, as hypothesised by Slobin (2004a, and elsewhere), accessibility of manner expressions in a given language leads to increased, and habitual, attention to that semantic domain. Over time, this in turn can lead to both greater lexical diversity of manner expressions, with new manner verbs being constantly added to the lexicon, and to the emergence of new construction types encoding manner of motion. In the case of English, this suggested line of development appears to account for the rise of patterns such as the way-construction (*the coach rumbled its way for miles*) discussed by Israel (1996), and the SEtoM Cxn discussed in the previous pages.
Beyond this, I also hope to have shown that lexicographic resources such as the evolving *OED3* and *HTOED* offer unparalleled access to large amounts of information on word histories, the histories of semantic fields and, ultimately, the histories of grammatical constructions. When the level of generalisation aimed for is the development of the English language generally rather than the history of any specific variety or register, the advantages of such resources far outweigh their drawbacks.

### Primary sources


Chadwyck-Healey. *Early American Fiction 1789-1875*.


Chadwyck-Healey. *English Drama* (late 13th c.-early 20th c.).


References


Fanego, Teresa 2012. Motion Events in English: The Emergence and Diachrony of Manner Salience from Old English to Late Modern English. *Folia Linguistica Historica*. 33, 29-85.


Slobin, Dan I. 2006b. Typology and Usage: Explorations of Motion Events Across Languages. Plenary lecture delivered at the *Fifth International Conference of the Spanish Cognitive Linguistics Association (AELCO)*, University of Murcia, 19-21 October 2006.


The use of *if* as a declarative complementizer in English: theoretical and empirical considerations

Abstract

This chapter approaches the study of the subordinator *if* in declarative complements of the type *It would be a good idea if you hired a bodyguard*. Like other connectives originally associated with adverbal subordination (e.g., *lest, as if, as though*) and with [+wh] complementation (e.g., *how*), over time *if* develops a subsidiary function as a declarative link. A number of syntactic and semantic criteria suggest the existence of gradience between (i) the conditional and the declarative complementizer use of *if*; and (ii) interrogative *if* and declarative *if*. Based on data from the *Helsinki Corpus* and *A Representative Corpus of Historical English Registers*, our study shows that the selection of declarative *if* in the history of English is typically associated with the presence of commentative predicates and non-assertive contexts. Moreover, *if*-complements are particularly common in subject function and show a lesser degree of integration in the sentence than prototypical subject clauses. Finally, data from *ARCHER 3.1* also suggest a close link between declarative complements introduced by *if* and informal style.

---

1 Authors’ email addresses: mjlopez.couso@usc.es and belen.mendez@usc.es; affiliation: U. of Santiago de Compostela. We gratefully acknowledge the financial support of the European Regional Development Fund and the following institutions: the Spanish Ministry of Economy and Competitiveness (grant FFI2011-26693-C02-01) and the Autonomous Government of Galicia (Directorate General for Scientific and Technological Promotion, grant CN2012/012).
1. Introduction

The extensive literature on finite declarative complementizers has paid considerable attention to the variation between the major complement-clause links *that* and zero, from both a synchronic perspective (cf. Elsness 1982, 1984; Tagliamonte/Smith 2005; Kaltenböck 2006; Kearns 2007; Torres Cacoullos/Walker 2008) and a diachronic one (cf. Warner 1982; Fanego 1990; Rissanen 1991; Finegan/Biber 1995; López-Couso 1996, among others). In addition to *that* and zero, declarative complements have occasionally been introduced at different stages in the history of the language by connectives typically associated with other types of subordinate clauses. The inventory of the items which can serve such a linking function in finite complementation includes, among others, the following: *but, how, lest, (as) if, (as) though, like, when, where, as, at, and because.*

Over recent years we have developed an interest in these so-called minor declarative complementizers, and have already examined the origin and development of some of these links (cf. *but* (*that*) (López-Couso/Méndez-Naya 1998), *if* and *though* (López-Couso/Méndez-Naya 2001), *lest* (López-Couso 2007), *as if, as though, and like* (López-Couso/Méndez-Naya 2012a, 2012b), and *how* (López-Couso/Méndez-Naya, in preparation)). The aim of this chapter is to contribute further to the understanding of this relatively neglected area of research in complementation studies, focusing on the declarative use of the originally adverbial link *if* from its earliest occurrences in Old English (OE) to the present day. The study thus complements our 2001 article, in which we traced the early history of *if* and *though* up to the Early Modern English (EModE) period. In the present chapter we will explore, among other issues, the following: (i) the relevance of structural factors such as predicate type and clause type; (ii) the degree of integration of *if*-complements in the sentence; and (iii) the potential influence of textual factors, such as text-type and degree of formality, on the selection of *if* as a declarative complementizer. We will use the two reference historical corpora covering the entire history of the English language, namely the Helsinki Corpus of
The use of *if* as a declarative complementizer in English

*English Texts (HC)* and *A Representative Corpus of Historical English Registers (ARCHER 3.1)*.

2. The subordinator *if*

From OE times to the present day, the subordinator *if* has normally been associated with adverbial clauses of condition, as in (1), and with indirect polarity questions, as in (2):

(1) **If** you don’t study, you won’t pass the exam.
(2a) I asked her *if* she wanted a drink.
(2b) I don’t know *if* she wants a drink.

The matrix in a prototypical conditional construction, like that in (1), expresses a state of affairs that is contingent on that of the dependent clause. In other words, the consequent state or event in the apodosis does not hold unless the condition stated in the protasis does. In turn, dependent interrogatives introduced by *if*, as in (2), are normally used to report queries corresponding to non-dependent *yes-no* questions; for instance, (2a) would correspond to a direct question such as the one in (3):

(3) I asked her: ‘Do you want a drink?’

Despite the existence of significant syntactic differences between the two types of constructions introduced by *if* in examples (1) and (2), a close look at these illustrative instances shows that there are important semantic resemblances between the two basic uses of the subordinator *if*: both conditional *if* and interrogative *if* “convey doubt about the truth value of the clause” (Quirk et al. 1985: §15.6, note a). Noonan (1985: 105) also notes that *if* is associated with non-actual or *irrealis* modality in both complementation and conditional subordination. Thus, the conditional construction in example (1) entertains two possibilities: one in which you pass the exam and one in which you do not. Similarly, there are two possible answers to the polarity question in
either she wants a drink or she does not. In other words, what seems to be at issue here is that in both cases we are dealing with non-assertive contexts.2

The semantic similarity between the conditionals and the indirect polarity questions described above explains the existence of examples such as (4), which are ambiguous between the two readings: “In the conditional interpretation I am asking you to tell me that you need help if you do, and in the interrogative interpretation to tell me whether or not you need help” (Huddleston/Pullum et al. 2002: 975):

(4) Let me know if you need any help. (from Huddleston/Pullum et al. 2002: 975)

The existence of this semantic connection between the conditional and the interrogative constructions under discussion here is not at all surprising (cf. Huddleston/Pullum et al. 2002: 970, 975), given that the English conditional subordinator if seems to be the source of the interrogative if (cf. Noonan 1985: 47; Traugott 1985: 291). Another example of the developmental path from conditional to interrogative is Romance si/se (cf. Harris 1986: 276). Note, however, that the reverse process, that is, from interrogative to conditional, is also attested crosslinguistically (cf. Traugott 1985: 291; Hopper/Traugott 2003: 186).

In addition to the two basic uses just described, the subordinator if can occasionally be found introducing finite declarative complement clauses, like the one here:

(5) It would be a good idea if you hired a bodyguard. (from McCawley 1988: 143)

Similar instances of this additional use of if are recorded since OE times. Illustrative examples from OE, Middle English (ME), EModE and Late Modern English (LModE) are given in (6a-d):

(6a) Lareow, ne oþingð hit ðe gif ic þus wer gecese? (HC, QO3_NI_FICT_APOLL) (from López-Couso/Méndez-Naya 2001: 96)

‘Teacher, doesn’t it displease you if I choose a man so?’

(6b) “If a blind man here vnderta / To lede anoþer blind alswa, / It es noght ferly if þai fall / Both in þe dike, for so þai sall”. (HC, QM3_IR_HOM_NHOM)

We follow here Quirk et al.’s (1985: §2.53) definition of non-assertiveness.
(6c) When he had obtain’d in England what he desir’d, no wonder if he sought the love of his conquer’d Subjects for the love of his own quiet, the maintainers of his wealth and state, for his own profit. (HC, QE3_NN_HIST_MILTON)

(6d) We are all well, I assure you; but you should not wonder if you find us rather grave. (ARCHER, 1839mart.f5b)

However, in spite of its continuous occurrence from the earliest stages of the language through to the present day, the declarative use of if has received very little attention in the literature on complementation. The infrequent occurrence of if as a declarative complementizer compared to its major uses as a conditional marker and as an interrogative complementizer may go a long way towards explaining this neglect (see §3 below). Nevertheless, passing mention of this declarative use of if can be found in Lakoff (1968: 69), McCawley (1988: 143), Mitchell (1985: §§1960, 3409-3410), Noonan (1985: 104-105), and the Oxford English Dictionary (OED) entry for though, where it is said that though can be replaced by if “in more or less weakened or modified sense [...] after negative or interrogative phrases with wonder, marvel, be sorry, care, etc.” (OED s.v. though II.4.a). The complement analysis for if-clauses of the type shown in (5) has, however, been rejected by authors such as Huddleston (1971: 177-178), given that they do not behave like standard cases of complementation (cf. §3.3 below). Although it may be admitted that the behaviour of the if-clauses in examples (5) and (6) does in fact depart from that of prototypical declarative complement clauses in certain respects, we believe that there are also convincing arguments for considering them as declarative complements. In what follows we will first establish a distinction between conditional, interrogative, and declarative uses of the subordinator if by considering a number of defining semantic and structural criteria which may help in disambiguation (cf. §§2.1, 2.2 and 2.3). We will then move on to the analysis of the history of if as a declarative complementizer as represented in the corpus material (cf. §3).

3 Mitchell (1985: §§3409-3410), in the sections devoted to “clauses with peaḥ (pe) for peér”, mentions in passing the complementizer use of gif in general comments of the kind “gif may also appear” or “here too we also find gif.”
2.1. Conditionals vs. declarative complements

As already noted, in a prototypical conditional construction like that in (1) above, the truth of the proposition in the main clause depends on the fulfilment of the condition stated in the subordinate clause; in such cases, therefore, *if* can be replaced by *in case* (cf. (1a)). By contrast, a reading in terms of ‘if/in case *P*, (then) *Q*’ makes no sense in an example like (5), where the idea of contingency is weakened or even nonexistent (cf. (5a-b)):

(1a) *In case you don’t study, you won’t pass the exam.*
(5a)  *If you hired a bodyguard, then it would be a good idea.*
(5b)  *It would be a good idea in case you hired a bodyguard.*

As opposed to conditional clauses, the *if*-clauses in examples (5) and (6) are subcategorized for by a complement-taking predicate (*be a good idea, of/pincean* (‘displease’), *be not ferly* (‘be no wonder’), *no wonder, not wonder*). They are therefore obligatory constituents in clause structure. Example (7) is particularly telling in this regard, since it contains two *if*-clauses: on the one hand, a conditional clause (*If God has endowed you with intellectual gifts*), which can be omitted without affecting the grammaticality of the sentence; and, on the other, a complement clause (*if you directly devote them to His service in the religious instruction*), whose omission leads to an ungrammatical sequence. It should be recalled here that obligatoriness is precisely one of the central criteria for distinguishing complements from adjuncts (cf., among others, Huddleston/Pullum et al. 2002: 219-228).

(7)  *If God has endowed you with intellectual gifts, it may be well if you directly devote them to His service in the religious instruction.* (ARCHER, 1857cair. h6b)

In addition to the obligatoriness criterion, the *if*-clauses in (5) and (6) behave like prototypical complements as regards two further criteria (cf. McCawley 1988: 143), namely pronominalization and pseudo-clefting. As regards pronominalization, the *if*-clauses in examples (5) and (6) pronominalize in the same way that complements do; cf. (5c) and (8). Concerning pseudo-clefting, they match the *what* of a pseudo-
cleft construction, as in (5d). Obviously enough, these two criteria do not apply to conditionals:

(5c)  This/that would be a good idea.
(8a)  It surprised me that he arrived so late.
(8b)  This/that surprised me.
(5d)  What would be a good idea would be if you hired a bodyguard.

Further evidence in favour of the complement analysis for these clauses is the fact that they can be replaced by unambiguous declarative complement clauses, both finite (introduced by either that or zero) and non-finite (infinitival and -ing clauses). Thus, for example, the infinitive construction in (5e) below can substitute for the if-clause in our earlier instance (5), with no apparent difference in meaning:

(5e)  It would be a good idea for you to hire a bodyguard.

Interchangeability between if and the major declarative complementizer that is particularly conspicuous in examples like (9), where the if-clause resumes a that-clause which has been truncated by intervening material:

(9)  Neither is it to be admired that Henry, who was a wise as well as a valiant prince, who claimed by succession, and was sensible that his title was not sound but was rightfully in Mortimer, who had married the heir of York; it was not to be admired, I say, if that great politician should be pleased to have the greatest wits of those times in his interests, and to be the trumpet of his praises (1668, John Dryden, Of Dramatic Poesy: An Essay 561/3-8; from López-Couso 1994: 345).

As we have seen, these criteria may be helpful for disambiguation in a good number of cases. It seems, however, that not all examples can so straightforwardly be ascribed to one or the other side of the boundary between conditionals and declarative complements. Indeterminacy between the two readings is evident in cases where the if-clause would be taken as a complement to an adjectival predicate, such as uneasy (OED s.v. uneasy a. ‘displeased, dissatisfied’) or sorry in examples
(10a-b). Although the if-clauses in such instances fulfil the structural criteria noted above for complements, they also allow the reading ‘if/in case P, (then) Q’ (e.g., I am sorry in case I look strange / If I look strange, I am sorry).

(10a) You would have been very uneasy if you had been a week without hearing from me. (ARCHER, 1764bswl.x4b)
(10b) (Lamorre) …Why do you look so strange?
     (Duke) I? Strange? […] I am sorry if I look strange. It may well be so. (ARCHER, 1943haml.d7b)

2.2. Interrogative vs. declarative complements

In addition to those cases of indeterminacy between conditional and complement clauses discussed in the preceding section, ambiguity may also arise within the domain of complementation itself. Declarative and interrogative if-clauses share a number of features. Being complements, the if-clauses in our earlier instances (2) and (5) behave in the same way with respect to the criteria of obligatoriness, pronominalization, and pseudo-clefting. However, they differ as regards clause type: while the clause dependent on the complement-taking predicate ask (2a) is interrogative, the one complementing the predicate be a good idea (5) is not, and as such does not allow the insertion of or not (cf. (11a) vs. (11b)). Notice also that, in such cases, if is not replaceable by the interrogative complementizer whether (cf. (11c) vs. (11d)):

(11a) I asked her if she wanted a drink or not.
(11b) *It would be a good idea if you hired a bodyguard or not.
(11c) I asked her whether she wanted a drink or not.
(11d) *It would be a good idea whether you hired a bodyguard or not.

Although the distinction between examples (2) and (5) is fairly straightforward, it seems that certain if-clauses are potentially ambiguous between an interrogative and a declarative reading. A par-

4 Other predicates of this kind include astonished, delighted, disappointed, glad, grateful, and surprised.
particularly interesting case is provided by the complement-taking predicate *doubt* in examples of the type shown in (12). In such instances, the distinction between the declarative and interrogative readings seems to be cancelled. On the one hand, the insertion of *or not* in these instances is not possible (cf. (12b)), while on the other hand *if* can be safely be replaced by *whether* without any apparent semantic difference (cf. (12c)). Extrapolating Huddleston/Pullum et al.’s (2002: 984) remarks on *doubt whether*-constructions to *doubt if*-clauses, example (12a) would involve a “mismatch between syntax and semantics: syntactically the complement is interrogative, whereas semantically it does not express a question”. In other words, the sequence in (12a) can be seen as equivalent to (12d):

(12a) Oh yes. I believe in that. And I grant you that in war my country has a perfect right to demand my life. I doubt *if it has the right to demand more*. (ARCHER, 1960ratt.d8b)
(12b) *I doubt *if it has the right to demand more or not.*
(12c) I doubt *whether it has the right to demand more.*
(12d) I doubt *that it has the right to demand more.*

2.3. *Gradience in if*-clauses

The evidence presented so far seems to suggest the existence of gradients between the three types of *if*-clauses discussed in the preceding sections. The proposed continuum between conditional *if*-clauses and declarative *if*-clauses is represented in Figure 1, explained below:

(i) The left endpoint on the scale would be occupied by those *if*-clauses which can only be interpreted as having conditional value: ‘*if/in case P, then Q*’. Prototypical instances of this kind are given in example (1) and the first *if*-clause in example (7).

(ii) Instances like those in examples (10a-b), in which the *if*-clause depends on an adjectival predicate, would be placed towards the declarative end of the scale. In such cases, the *if*-clause fulfils the criteria for complementhood, namely obligatoriness, pronominalization, pseudo-clefting, and replacement by an unambiguous declarative complement, but also allows a reading in
terms of ‘if/in case $P$, (then) $Q$’. Such instances would therefore represent an intermediate point along the gradient.

(iii) Finally, at the declarative end of the scale we recognize examples such as (5), (6a-d), or the second clause in example (7), which fulfil the four criteria for complements. In such cases, in which a conditional reading makes no sense at all, *if* can therefore be seen as an equivalent (or near-equivalent) of the default declarative complementizer *that*.

![Figure 1. The continuum between conditional *if* and declarative *if*.](image)

In turn, the continuum between the interrogative complementizer *if* and the declarative complementizer *if* is shown in Figure 2, explained below:

(i) In this second continuum, *if*-clauses which can only be interpreted as [+wh]-complements occupy the left endpoint on the scale. Prototypical instances of this kind show a verb of asking in the matrix, such as *ask* in (2a) above, allow the insertion of *or not*, and also allow the replacement of *if* by *whether*.

(ii) Instances like (2b) above, which depend on a predicate other than one of asking, but fulfil criteria (b) and (c) for interrogative clauses, would occupy a middle position in the continuum.
(iii) In turn, if-clauses dependent on the complement-taking predicate doubt, as in example (12a), would occupy an intermediate point on the scale: on the one hand, the possibility of replacing if by whether makes them close to being interrogative complements; on the other hand, the fact that such instances do not allow the insertion of or not, while admitting replacement by an unambiguous declarative complement, would indicate that such if-clauses are in fact more declarative-like than interrogative-like.

(iv) Once again, the right endpoint of the scale would be occupied by those examples in which the if-clause is clearly declarative, and in which if is therefore an equivalent (or near-equivalent) of the default declarative complementizer that.

| Interrogative if | ← | (i) Indisputable instances of interrogative if: criteria (a)-(c). | (ii) If-clauses dependent on verbs other than of asking: criteria (b)-(c). | (iii) If-clauses dependent on the predicate doubt: criteria (c)-(d). | (iv) Clear instances of declarative if on structural and semantic grounds: criterion (d). | → | Declarative if |

Criteria:

(a) Predicate of asking
(b) Insertion of or not
(c) Replacement by whether
(d) Replacement by unambiguous declarative complements

Figure 2. The continuum between interrogative if and declarative if.

3. The data

The data for the present study are drawn from the HC and the six sub-periods representative of 18th, 19th and 20th century British English in ARCHER 3.1. A total of 2,646,188 words have been analysed.
Table 1 shows the number of words analysed for the present study and their distribution across time periods:

<table>
<thead>
<tr>
<th>PERIOD</th>
<th>TIME SPAN</th>
<th>CORPUS</th>
<th>NO. OF WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1150</td>
<td>HC</td>
<td>413,250</td>
</tr>
<tr>
<td></td>
<td>1150-1500</td>
<td>HC</td>
<td>608,570</td>
</tr>
<tr>
<td></td>
<td>1500-1710</td>
<td>HC</td>
<td>551,000</td>
</tr>
<tr>
<td></td>
<td>1700-1900</td>
<td>ARCHER</td>
<td>718,220</td>
</tr>
<tr>
<td></td>
<td>1900-1990</td>
<td>ARCHER</td>
<td>355,148</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td></td>
<td>2,646,188</td>
</tr>
</tbody>
</table>

Table 1. Distribution of words in the working corpus.

3.1. Function and predicate

The analysis of the two corpora yielded a total of 8,105 examples of *if*. Careful examination of these examples revealed that only 90 instances (1.1%) fulfilled the different criteria established above for *if*-declarative complements, which testifies to the marginality of the function under investigation here among the possible functions of the conjunction *if*. Table 2 shows the distribution of the relevant instances per period and function:

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>LModE-BrE</th>
<th>20th c.-BrE</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>28</td>
<td>13</td>
<td>25</td>
<td>8</td>
<td>75</td>
</tr>
<tr>
<td>OBJECT</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1</td>
<td>28</td>
<td>17</td>
<td>32</td>
<td>12</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 2. *If*-declarative complements in the *HC* and *ARCHER* 3.1 (raw figures and normalized frequencies per 100,000 words).

As Table 2 shows, the vast majority of relevant examples (83.3%) are cases in which the *if*-clause is amenable to a subject analysis, as in our earlier instances (6a-c), while object clauses, as in (6d), represent only 16.7% of the total. This is indeed one of the respects in which *if*-complements differ starkly from prototypical declarative complement clauses introduced by *that* or zero, which show, both for Present-day English (PDE) and for earlier stages of the language, a much higher proportion of the object function (around 70%; cf. Elsness 1981:...
The use of *if* as a declarative complementizer in English


The predicates typically introducing subject *if*-clauses take the shape of a copula verb + NP (38 ex.) or an AdjP (30 ex.),\(^5\) as seen in examples (6b-c). Some illustrative predicates of these types are: be error, be no fault, be great sin (NP), be bad, be good, be strange (AdjP). In the case of object clauses, the relevant predicates are bear/ not bear (2), marvel, not wonder (7), prefer, not resent, take it for a great favour, take it ill, think it enough. Note that predicates meaning ‘wonder’ like (be) (no/little/not much) wonder (11), be marvel (2), not wonder (7) clearly prevail in the data.

The matrix predicates selecting *if*-declarative complements in the corpus typically belong to the group of commentatives in Noonan’s taxonomy of complement-taking predicates (1985: 117), i.e., those expressing emotional reactions (e.g., wonder, marvel, resent, etc.), evaluations, or judgements (be better, be well, be strange, be (an) error). Interestingly enough, commentatives frequently surface as one-place predicates, which could explain (at least in part) the aforementioned preference for *if*-declarative clauses to realize the subject function.

Such a restriction to a particular class of predicates seems to be a characteristic of so-called minor complementizers. Thus, for example, the declarative complementizer *but* normally occurs with predicates of negative import, especially those denoting doubt (López-Couso/Méndez-Naya 1998), while the complementizer *lest* seems to be confined to clauses dependent on predicates of fearing (cf. López-Couso 2007: 18-19), i.e., those expressing “an attitude of fear or concern that the complement proposition will be or has been realized” (Noonan 1985: 119). In turn, the comparative complementizers *as if*, *as though*, and *like* are normally associated with propositional attitude predicates (e.g., *look, seem, sound*) expressing a medium degree of speaker commitment towards the truth of the proposition (cf. López-Couso/Méndez-Naya 2012a, 2012b).

---

\(^5\) There are seven examples that do not conform to this pattern: be misconstrued into pride, be to me, reflect shame, set ne at ease, serve right, icwemen, and ne of pincean.
3.2. If-complements and non-assertiveness

In López-Couso/Méndez-Naya (2001) we showed that the occurrence of *if*- and *though*-complements also proves to be very closely related to the presence of a non-assertive context in the matrix clause. As far as *though* is concerned, the few references to this link in the literature point out that it is typically found in two types of non-assertive environments, namely interrogative and negative matrices (*OED* s.v. *though* II.4.a; *MED* s.v. *though* 6; Mitchell 1985: §3407-3015; López-Couso/Méndez-Naya 2001: 100). In the case of *if*, by contrast, affirmative matrices prevail (cf. López-Couso/Méndez-Naya 2001: 102). However, in such cases non-assertiveness is achieved by other means. Thus, for example, the predicate may itself be inherently negative (e.g., *fault, shame, sin, strange*, etc.), as in example (13a) featuring the predicate *be shame*. In other cases, the VP in the matrix clause is modally marked, by means of either a subjunctive VP, *were* in (13b), or a modal periphrasis, *would be* in (13c) (cf. Quirk et al. 1985: §10.61). As the examples show, the original positive implication of the predicates *be good* and *be a real comfort* are cancelled by the presence of the subjunctive and the modal in the VP, so that they become comparable to negated matrices:

(13a) And shame it is, *if* a prest take keep, / a shiten sheperde and a clene sheep. (*HC, QM3_NI_FICT_CTPROL*)
(13b) And that they may not squander away their time in idleness, it were good *if* they were put to a Writing-Schoole, (*HC, QE3_IS_EDUC_HOOLE*)
(13c) [I]t would be a real comfort to me *if* you would make me feel we belonged to each other. (*ARCHER, 1893pine.d6b*)

The close connection of *if*-declarative complements and non-assertiveness in the two corpora used for the present study becomes obvious from the data in Table 3:

<table>
<thead>
<tr>
<th></th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>LModE</th>
<th>20th c.</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSERTIVE MATRIX</td>
<td>-</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>NON-ASSERTIVE MATRIX</td>
<td>1</td>
<td>26</td>
<td>12</td>
<td>30</td>
<td>11</td>
<td>80</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1</td>
<td>28</td>
<td>17</td>
<td>32</td>
<td>12</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 3. *If*-declarative complements and assertiveness.
The use of *if* as a declarative complementizer in English

The existence of such an intimate relation between *if*-declarative complements and non-assertiveness is not at all surprising if we consider that, as mentioned in §2 above, the origin of the *if*-complementizer (both interrogative and declarative) is found in its conditional use, which is also closely connected with non-assertiveness. As Noonan (1985: 104) observes, “when a form functions as a complementizer and something else, its meaning outside the complement system will likely be related to its use in complementation”.

Although there seems to be a very strong tendency for *if*-complements to appear in connection with non-assertive matrices, we have recorded 10 examples in the corpus which contain an assertive matrix clause. In all such instances, however, the content of the *if*-clause is taken to convey a putative situation; in five cases, the *if*-complement is itself marked for the subjunctive, as shown in example (14) with the form *doo* (vs. indicative *doest*):

(14) and with an earnest study resorteth often to looke vpon this booke I maruell *if* he doo not fynde here in, great cause and earnest matter of dread. (HC, QE1_IR_SERM_FISHER)

Even though non-assertive contexts favour the selection of the minor complementizer *if*, the major complement-clause links *that* and zero may also occur with commentative predicates in non-assertive environments. Such instances are, however, rather exceptional in the corpus. Examples (15a-b) illustrate the use of the two major links after the matrix predicate (*be*) *no wonder*:

(15a) No wonder Ø *I hadn’t recognized him at first.* (ARCHER, 1923wode.f7b)
(15b) [...] it is therefore no Wonder *that so many Trees miscarry in Planting, when there are no Branches left on the Head.* (ARCHER, 1724fair.s3b)

It must be noted, however, that the declarative complementizers *that* and *if* do not always seem to be interchangeable in PDE. Huddleston/Pullum et al. (2002: 1254) report the impossibility of using the complementizer *that* in an example like (16a), where the conditional implication of the VP *would be* seems to be inconsistent with the presence of a factive complement (cf. Quirk et al. 1985: §16.72 note b). Note that this example illustrates one of the conditioning environ-
ments for *if*-complements, namely the presence of a non-assertive context in the matrix (see above). Example (16a) contrasts with that in (16b), where the complementizer *that* is perfectly grammatical:

(16a) It would be good *if / *that you invited them both.
(16b) It was good *that you invited them both. (from Huddleston/Pullum et al. 2002: 1254)

3.3. If-complements and clause integration

As discussed in §3.1, most of the *if*-declarative clauses recorded in the *HC* and *ARCHER 3.1* would allow a subject analysis (83.3%). However, some of the complements we have classified here as subject clauses are indeed non-prototypical subject clauses, since they show a somewhat peculiar behaviour both in terms of position and the presence of an anticipator. Table 4 shows the distribution of *if*-subject clauses in the corpus according to the position they occupy in the sentence. As can be seen, 72% out of the 75 examples of subject clauses identified in the corpus are extraposed, that is, they occur after the matrix verb, the canonical subject position being occupied by an anticipatory pronoun (cf. instances (6a-b); (15b)). Note that pre-verbal position has been the unmarked position for subject clauses from OE to the present day (cf. López-Couso 1994: 108-111), so that extraposed subject clauses can be regarded as the statistical prototype in English (Mair 1990: 37). It must be noted, however, that there is a remarkable difference between *if*- and *that*-complements as regards the proportion of preverbal clauses. Taking the corpus as a whole, the percentage of preverbal *if*-subject clauses reaches 28%, which is much higher than the average obtained for pre-verbal *that*-subject clauses in different synchronic and diachronic corpus-based studies on clausal complementation, at around 10% (cf., among others, Huddleston 1971: 176, 192; Ellegård 1978: 32; Fanego 1990: 9; Mair 1990: 22).

---

6 As Mair (1990: 37) puts it, the postverbal structure is “the statistical prototype and hence a neutral or unmarked form which is suitable for both given and new information, for very simple clauses and for long and complex ones”.

The use of *if* as a declarative complementizer in English

<table>
<thead>
<tr>
<th></th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>LModE</th>
<th>20th c.</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PREVERBAL</strong></td>
<td>-</td>
<td>11</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NF 1.8</td>
<td>NF 0.3</td>
<td>NF 0.8</td>
<td>NF 0.6</td>
<td>NF 0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.3%</td>
<td>15.4%</td>
<td>24%</td>
<td>25%</td>
<td>28%</td>
</tr>
<tr>
<td><strong>POSTVERBAL</strong></td>
<td>1</td>
<td>17</td>
<td>11</td>
<td>19</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NF 0.2</td>
<td>NF 2.8</td>
<td>NF 2.6</td>
<td>NF 1.7</td>
<td>NF 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>60.7%</td>
<td>76%</td>
<td>75%</td>
<td>72%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>1</td>
<td>28</td>
<td>13</td>
<td>25</td>
<td>8</td>
<td>75</td>
</tr>
</tbody>
</table>

Table 4. Position of *if*-subject clauses in the corpus (raw figures, normalized frequencies per 100,000 words, and percentages).

In addition to extraposed *if*-subject clauses, the corpus yields examples of preverbal *if*-clauses which are also susceptible to a subject analysis. Illustrative examples are given in (17a-c). The *if*-clauses in these instances differ from prototypical preverbal clausal subjects, such as the *that*-clause in (18), in two respects: on the one hand, the *if*-clause is immediately followed by a resumptive pronoun (*it* in most cases); on the other, some of these examples (cf. (17b-c)) show a punctuation mark which seems to suggest the existence of an intonational boundary between the main clause and its complement:

(17a) Also taketh a lytill bawme with the poynct of a knyf & touche it to the fuyr & *sif* *it* brenne *it* is a gode signe. *(HC, QM3_NI_TRAV_MAND)*

(17b) You had better go to bed, however; *if* *it* is known *that* *you* visit me here, *it* will be bad for us both. *(ARCHER, 1778reev.4fb)*

(17c) They would, I am sure, be the last to claim such an achievement, but whatever balance they strike *if* *I* could discover and live *it*, *it* would be more than enough for me. *(ARCHER, 1981long.j8b)*

(18) *That* he didn’t pass the exam was rather surprising.

These two features, along with the fact that *it* can be replaced by the demonstratives *this* or *that*, are the arguments adduced by Huddleston (1971: 177-178) to cast doubt on the status of *if* declarative clauses as complements in PDE (cf. §2 above). For Huddleston, the subject of

---

7 It must be noted, however, that punctuation in earlier stages of the language was far less regulated than nowadays and should therefore be viewed with caution.

8 Up to EModE, however, *this* and *that* could anticipate *that*-clauses, which weakens the argument for earlier periods of English.
the construction is not the *if*-clause but rather the pronoun *it*, which implies that *if* “is not the complementizer, but the ordinary conditional conjunction” (1971: 178).

A possible analysis for such clauses is one in terms of topicalized complements resumed by a pronoun (López-Couso/Méndez-Naya 2001; cf. Pérez-Guerra 1999: 138 for a similar analysis). Examples comparable to those in (17a-c), containing *that*- or infinitival clauses, are not difficult to find. Consider (19a-b):

(19a) That I haue tane away this old mans Daughter./ It is most true. (*Othello* I,iii, 78; from Barber 1976: 284; Visser 1963-1973: §73.a.)
(19b) [10 sec. untranscribable due to extraneous noise] well to / start off at half cock # /you know#. /it’s stupid# (S.1.2a.19; from Mair 1990: 56)

Clauses such as (19b) are, in Mair’s (1990: 56) terms, ‘semi-autonomous clauses’, which can be regarded as “the informal equivalent of non-extraposed subject clauses” (1990: 56). Such structures could be said to be placed between the two cluster points of hypotaxis and subordination in the continuum of clause combining proposed by Hopper/Traugott (2003: 176-184) in terms of parataxis > hypotaxis > subordination (where > indicates a lower degree of integration in a complex sentence) (cf. López-Couso/Méndez-Naya 2001: 99).

Comparison of the relative proportions of preverbal vs. postverbal *if*-clauses across periods shows a marked decrease of the preverbal position from ME onwards, which could be taken as an indication of a growing degree of clausal integration for *if*-complements over time.

Topicalized object *if*-clauses are also found in the data. The corpus yielded two examples of this kind, which are reproduced as (20a-b) below. The vast majority of object *if*-clauses, however, are found in postverbal position.

(20a) *If Lucas and I are parted I can’t bear it! I can’t bear it! (ARCHER, 1847carl.f5b)
(20b) “Why, in troth, Rody, to tell you the truth *if she could give it to me at nay other time an’ place,* I’d prefer it. (ARCHER, 1847carl.f5b)
3.4. If-complements and text-type

In previous sections we have paid attention to different structural and semantic features which are relevant to the characterization of declarative complements introduced by *if* in the history of English. Further interesting insights can be obtained from the consideration of textual factors, such as the correlation between *if*-complements and text-type, which we will examine here using data from *ARCHER 3.1*.

The basic distinction established by the compilers of *ARCHER 3.1* is that between speech-based registers, which comprise Fictional conversation, Drama, and Sermons-Homilies, and Written registers, which include the remaining text-types (Journal-Diaries, Letters, Fiction, News, Legal Opinion, Medicine, and Science). In both groups, both formal and informal writing are represented. Table 5 further down provides the figures for *if*-declarative complements according to the dichotomy ‘speech-based’ vs. ‘written registers’ in the *ARCHER 3.1* material.

As Table 5 clearly shows, *if*-complements exhibit a strong preference for speech-based registers, with declarative *if* occurring four times more frequently than in the written categories (NF 8.7 vs. 2.1). Further evidence of the close connection of the declarative complementizer *if* and speechlikeness is provided by its higher frequency in those text-types illustrating the most informal end of both the speech-based group (fictional conversation and drama) and the written registers (letters and journals-diaries). Interestingly, the intimate connection between declarative *if* and informal style finds a parallel in the behaviour of other minor complementizers, such as *as if, as though*, and *like*, which also tend to occur more readily in informal language (cf. López-Couso/Méndez-Naya 2012a).
SPEECH-BASED OCCURRENCES WRITTEN OCCURRENCES
REGISTERS registers
Fictional conversation 10 (NF 9.4) Journal-Diaries 5 (NF 3.8)
Drama 15 (NF 10) Letters 4 (NF 5.6)
Fiction prose 3 (NF 1.8) News 1 (NF 0.7)
Sermons-Homilies 3 (NF 4.6) Medicine 2 (NF 1.6)
Science 1 (NF 0.8)
TOTAL 28 TOTAL 16
NF 8.7 NF 2.1

Table 5. *If*-declarative complements in the different register categories in *ARCHER 3.1* (raw figures and normalized frequencies per 100,000 words).

4. Conclusions

In this chapter we have studied the declarative use of *if*, a subordinator normally associated with adverbial clauses of condition and with dependent questions. Although declarative *if* has been attested from OE onwards, its use over the course of the recorded history of English is certainly a very marginal one.

We have proposed a number of semantic and structural criteria in order to establish a distinction between conditionals and dependent questions introduced by *if*, on the one hand, and cases of declarative *if*, on the other. It should come as no surprise that the boundaries between these categories are not always well-defined, since the three constructions share a close connection with non-assertiveness.

The survey of the data has revealed that the most common function for *if*-declarative complements is that of subject. The most frequent type of matrix is a non-assertive one, containing a copula verb + NP or AdjP and denoting a judgement or an emotional reaction (commentative predicates). A good number of these subject clauses are topicalized and resumed by a recapitulatory pronoun and are therefore less integrated in the sentence than prototypical subject clauses.

As far as the relevance of text-type is concerned, the data from *ARCHER 3.1* clearly suggest the existence of a particularly strong link
The use of if as a declarative complementizer in English

between if-complements with speech-based registers and informal style.

References


López-Couso, María José / Méndez-Naya, Belén In preparation. Tracking Down the History of Declarative How: Evidence from the Helsinki Corpus and ARCHER.


On English historical corpora, with notes on the development of adverbial connectives¹

Abstract

Like the attitude towards the role of the history of English for the proper understanding of contemporary English, computerized corpora have evolved to a point where they are essential for diachronic studies for their exhaustive representation and coverage of language features and also for the ease of data retrieval.

This paper reviews some of the major corpora used for the analysis of change in the history of English. It illustrates the features of these corpora by a brief survey of the development of the adverbial connectives *nemne/nympe* ‘except’, *according to*, and *concerning* in Old, Middle, Early Modern and Late Modern English. By doing so, these adverbial connectives also become the subject matter of the paper. The emphasis is on variation, with special attention to the influence of grammaticalization and to the role of language contact in the long-term developments of adverbial connectives.

1. Introduction

In the last few decades, computerised corpora have opened new vistas for the study of the history of English, particularly for long-diachrony surveys covering the periods from the earliest stages of the language to Present-day English (PDE) (see, e.g., Lüdeling/Kytö 2008). Empirical, text-based analysis became more solid and reliable as the results could be more easily verified or falsified. The most concrete advantage was of course the enormous saving of time in collecting evidence

¹ Author’s email address: mrissane@mappi.helsinki.fi; affiliation: U. of Helsinki.
in studies of long diachrony as well as past synchrony, which made it possible to handle much larger entities of text than earlier. Furthermore, the fact that we can only really understand the essence of today’s language if we understand its earlier forms and developments is now more obvious and easier to accept.

The new vistas opened by historical corpora also meant a decisive advance for the variationist approach to the study of linguistic change. The influence of extra-linguistic factors – sociolinguistic, regional or genre-based – on the choice of variant forms and expressions could be observed and compared more readily than earlier (see, e.g., Rissanen 2012). Variationist analysis based on rich corpus-based evidence also offers a solid basis for the new important approach in the study of linguistic, mainly semantic, change: the Invited Inferencing Theory of Semantic Change (IITSC), introduced in detail in Traugott/Dasher (2002). This approach emphasizes the role of the communicative aspect in language change and development. “The speaker/writer evokes implicatures and invites the addressee/reader to infer them” (Traugott/Dasher 2002: 5). In other words, when the speaker or writer uses a word or phrase in a novel way, the reader most naturally associates it with the other uses of the same word or phrase – the uses he or she is already familiar with. And when this new form becomes part and parcel of the hearer’s stock of expressions, a new variant is born.

The obvious advantages of corpus-based research of the history of English have created a continuous demand of new corpora, larger, more focused and more sophisticated in annotation. The Helsinki Corpus of English Texts (HC) (Rissanen/Kytö 1991; see also Rissanen et al. 1993), the first long-diachrony English general corpus, celebrated its twentieth anniversary in the autumn 2011. At the moment, the Corpus Research Database (CoRD),\(^2\) lists more than 50 historical English corpora, both general or multi-purpose and special or focused.

---

\(^2\) CoRD is maintained and developed by the Research Unit for Variation, Contacts and Change in English at the University of Helsinki, <http://www.helsinki.fi/varieng/CoRD/index.html>. At the moment, there are more than 50 entries in this database. CoRD does not give access to the corpora, only background information. This information includes the general character of the corpus and links to the manuals and compilers’ homepages.
The text material of the HC covers the first millennium of English writings, from the 8th to the beginning of the 18th century. It is divided into periods and sub-periods and is loosely structured, among other things, by genre, dialect and the sociolinguistic characteristics of the authors. There is also a new development of this corpus, the TEI XML version. The annotated Penn-Helsinki corpus family (Middle English (ME) and Early Modern English (EModE)) follows the same structural principles as the HC. The Penn Parsed Corpus of Modern British English (PPCMBE) (Kroch et al. 2010) covers the 18th and 19th centuries and the beginning of the 20th. There are also a number of other valuable period corpora: the Dictionary of Old English Corpus (DOEC), the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE), the Innsbruck Corpus of Middle English Prose (ICoMEp), and the Corpus of Middle English Prose and Verse (CMPEV). Two general corpora include texts dating from the 18th, 19th and 20th century: A Representative Corpus of Historical English Registers (ARCHER) and the Corpus of Late Modern English Texts (CLMET).

The most important sociolinguistically focused historical corpora are included in the Corpus of Early English Correspondence (CEEC) family. Of regional corpora, the Seville Corpus of Northern English (SCONE) covers the period from Early Old English to the end of the 16th century. The Helsinki Corpus of Older Scots (HCOS), the Helsinki Corpus of Scottish Correspondence, A Corpus of Irish English, and the gigantic Corpus of Historical American English (COHA) illustrate the history of important regional varieties. The number of genre-based corpora is large: the Corpus of Early English Medical Writing (CEEM), the Lampeter Corpus of Early Modern English Tracts, Zurich English Newspaper Corpus (ZEN) and the Corpus of English Dialogues (CED) are some of the most frequently used.

The purpose of the present paper is to briefly introduce some of the historical corpora mentioned in the preceding paragraphs, particularly those which are most useful for the analysis of morphosyntactic and lexical change in the long diachrony of the English language. Small, structured and focused corpora will be compared with large but less flexible corpora. The amount and quality of information given by these corpora will be illustrated by a brief survey of the development
of three adverbial connectives in Old English (OE), ME, EModE and Late Modern English (LModE). These items are nemne/nympe ‘except’, according to, and concerning.

My general approach applied to the analysis and description of the history of English is variationist, which, as mentioned above, is supported by information yielded by corpora. In the study of variation, language is first and foremost seen as a means of communication, oral and written, and the development and change in language should be understood and explained through variation which appears both in spoken and in written communication.

It should be emphasized, however, that corpus-based study of variation does not diminish the value of more theoretical approaches to historical linguistics. Generalisations and theoretical considerations add to the value of data-based results, while surveys of data are essential to trigger and support theoretical discussions. We should also keep in mind that in the study of linguistic variation the mere presentation of corpus-based quantitative results is not sufficient as such. The conclusions drawn from quantification through careful and insightful analysis are essential. “Research begins where counting ends” has always been the slogan at our research unit VARIENG in Helsinki.

Adverbial connectives offer a useful group of linguistic items to illustrate long-term developments of English grammatical vocabulary. This group of connectives consists of adverbs, prepositions and subordinators indicating certain relations between concepts and proposals, such as time, place, condition, concession, comparison, purpose and result. In historical studies it is impractical or even impossible to keep prepositions and subordinators separate: the function of the connective after in I’m going home after my lecture, I’m going home after giving my lecture, and I’m going home after I have given my lecture is more or less identical. In these three constructions after connects two actions and indicates their temporal relationship.3

There are two more aspects which make adverbial connectives a rewarding topic for the analysis of variation and change: one is gram-

3 It is worth noting that Huddleston/Pullum et al. (2002) define adverbial subordinators as prepositions governing a clause instead of a noun phrase.
mationalization and the other is early contact with other languages, mainly of course with French and Latin, and, to some extent, Old Norse (ON). The role of grammaticalization is obvious as the connectives typically belong to grammatical vocabulary. The often-quoted brief definition in Hopper/Traugott is adequate: grammaticalization is “the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions, and, once grammaticalized, continue to develop new grammatical functions” (2003: xv). Contact is essential in any study of ME vocabulary, and this also concerns the invasion of new connectives.

As to earlier studies of adverbial connectives, the most important is Bernd Kortmann’s epoch-making study Adverbial subordination (1997). Ursula Lenker (2010) has most thoroughly and competently covered the history of adverbial connectors, i.e., linking adverbs. As to the history of individual adverbial connective items, the studies of Rafał Molencki (e.g., 2012) and his group are of importance. See also Meurman-Solin/Lenker (2011) and the list of references at the end of this article.

2. Old English

As mentioned above, there are at least four important corpora including OE texts: the Dictionary of Old English Corpus (DOEC), the Helsinki Corpus of English Texts (HC), the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE) and the Seville Corpus of Northern English (SCONE). The SCONE includes not only OE, but also ME and EModE texts; the YCOE is based on a large number of OE prose texts. Both corpora are provided with word-class and syntactic annotation.

In the following I will introduce the most recent English historical corpus version, the TEI XML Version of the Helsinki Corpus, and compare its OE part with the DOEC. The DOEC includes practically all OE texts, some 3.5 million words, and is thus indispensable
for all extensive surveys of OE grammar and lexis. Because of its size, there are, however, some less practical features in its use. Unlike the HC, it is not structured by date, dialectal features or genres or types of text – essential information for all variationist study of the history of English. The OE part of the HC consists of c. 413,000 words and each text sample is preceded by a list of parameters describing the sample (Table 1; see also Kytö 1996).

| N CURA PASTORALIS = Name of text |
| A ALFRED = Author |
| C O2 = Sub-period |
| O 850-950 = Date of original |
| M 850-950 = Date of MS |
| K CONTEMP = Original & MS contemporary |
| D WS = Dialect (West Saxon) |
| V PROSE = Prose or verse |
| T REL TREAT = Genre (Religious treatise) |
| G TRANSL = Translated |
| F LATIN = Source language |
| Z INSTR REL = Text category (Religious instruction) |

Table 1. Set of parameters included in an OE text sample (King Alfred’s West-Saxon Version of Gregory’s Pastoral Care).

The TEI XML Version of the Helsinki Corpus of English Texts was publicized at the end of September 2011, in connection with the Helsinki Corpus Festival. This new version, still under development, will make it possible for the user to make searches according to the parameters shown in Table 1. I will compare the occurrences of the adverbial connective (preposition and subordinator) nemne / nympe ‘except’ in this new version and the DOEC. This connective is used by the side of butan ‘except’ and occurs in Anglian texts. It disappears at the end of the OE period. See examples (1)-(3).
(1) hine middangeardes þingum to þon ongyrede & genacodade, þæt he eal forlet, þa ðe he hæfte, nemne his anfealdne gegyrelan (Bede 264 BEDEHE HC)\(^4\) ‘he forsook worldly things to the extent that he let go all he had except his simple garment’

(2) þæt nænig mon ingan mehte & eft gesund æfter þon beon nymþe he mid asegendnissem ineode in þæt scræf. (Alexander’s Letter 36 TRAV ALEX HC) ‘no man could go in and after that be unhurt unless he went to the cave with an offering’

(3) [...] þæt ðum hringsele hrinan ne moste gumena ænig, nefne god sylfa, sigora soðcyning, sealde þam ðe he wolde (Beowulf 94 BEOW HC) ‘that no man could enter the ring-hall, unless God himself, the king of victories, gave permission to whom he chose’

Nymþe also occurs in the forms nefne, naefne, nemne, næmne, nymne, nemþe, nimþe, nymþe. Thorn (þ) and eth (ð) vary. Thus there are 13 spelling variants. This means 13 \(DOEC\) searches as the search program does not allow wild cards. The TEI XML HC version should give all the spellings at a single search.

A survey of the parameter list (Table 1) gives more interesting information based on the TEI XML HC version. A search based on various OE dialects as listed in Table 2 indicates that nymþe (in various spellings) does not occur in texts which are defined as purely West Saxon, and it occurs only once in a purely Kentish text while it occurs 13 times in texts that are defined as Anglian or Northumbrian, all except one in the Rushworth Gospels. The remaining examples occur in texts with mixed dialectal features. These frequencies point to non-West-Saxon character of nymþe, but are of course too low for definite conclusions. Another interesting observation is that out of the 56 HC examples, no less than 35 can be found in poetical texts (twelve in Beowulf).

\(^4\) The abbreviated title of the corpus from which the example is taken is indicated at the end of the reference line. The capitalized abbreviations in the Helsinki Corpus reference line refer to the texts listed in Kytö (1996: 169-230).
Table 2. ‘Dialect’ Reference Code Values in TEI XML HC.

A-X (Anglian dialect of Old English with unknown features)
AM (Anglian Mercian dialect of Old English)
AM-X (Anglian Mercian dialect of Old English with unknown features)
AN (Anglian Northumbrian dialect of Old English)
K (Kentish dialect of Old English)
K-X (Kentish dialect of Old English with unknown features)
WS (West Saxon dialect of Old English)
WS-K (West Saxon dialect of Old English with Kentish features)
WS-A (West Saxon dialect of Old English with Anglian features)
WS-AM (West Saxon dialect of Old English with Anglian Mercian features)
WS-X (West Saxon dialect of Old English with unknown features)

To confirm the results based on the HC, the 261 instances from the DOEC are of importance. This is a somewhat time-consuming task as every spelling variant must be searched separately and the text samples cannot be automatically grouped by dialect or other parameters. But this is of course a minor disadvantage in comparison to the amount of evidence. The great advantage of the DOEC is that it includes as complete a selection of OE nymþe forms as possible. All corpora – even multi-purpose corpora – are designed for certain types of research corresponding to the aims and visions of their compilers. The purpose of the DOEC is to give as much evidence as extant OE texts contain, while the purpose of the HC compilers was to offer the scholars a selection of texts and text samples that would make the comparison of texts of different kinds, defined by a set of parameters, as easy as possible and in this way enhance the variationist approach to linguistic change.

In searches and analyses focusing on later periods of English, the new version of HC is also useful, as the number and significance of the textual parameters increase. This version does not, however, at this stage, include grammatical tagging or parsing. For that kind of information on OE texts, the YCOE and the SCONE can be used.
3. Middle English

There is radical change in the number, variety and sources of adverbial subordinators in the course of the history of English, from the earliest stages of the language to the present day. In OE, the typical adverbial subordinators were phrases of the type preposition + demonstrative pronoun (oblique form) + an optional subordination marker þe or þæt: for þam þe, mid þy þe, and so on (examples (4)-(5)).

(4) he nyste hwæt þæs sōþes wæs, for þæm he hit self ne geseah (Oththere 17 OTHHR2 HC)
   ‘He did not know what the truth was because he did not see it himself.’

(5) ic ondræde me for ðam ðe ic com nacod, (Hept. Gen. 3 AELFOLD HC)
   ‘I am afraid because I am naked.’

There were of course simple subordinators, such as gif ‘if’, þeah ‘though’, and some subordinators were formed with an adverb or a noun, such as þy læs þe ‘the less that’, ‘lest’ or þa hwile þe ‘at that time that’, ‘while’. It is easy to see, however, that adverbial subordinate linking was still under development, native resources were used, and the grammaticalization of the items was incomplete.

The change from OE to ME was, indeed, dramatic. There was a flood of new adverbial connectives. Most of them were borrowed from French or Latin; a few from ON. Some were loan translations. This increase is understandable: new society and culture, new types of literature, both native and translated, new forms of laws, statutes and documents, all these markers of a new historical period encouraged and demanded new ways of linking ideas and propositions and indicating hierarchies between them.

Adverbial connectives borrowed from French and Latin, or loan translations, include, among others, because, in case, provided, save, except, in despite of, according to, outtaken/outnome, unless, notwithstanding, and all be it (that). It is worth noting that many of these new forms were based on verbs.
In the detailed study of the emergence and establishment of the new connectives ME corpora are indispensable. The half-a-million-word ME part of the \textit{HC} gives a good starting point. But, like in OE, the \textit{HC} results should be supplemented from other, much larger corpora.

At the University of Pennsylvania, three English period corpora have been produced: ME, EModE and LModE. The ME corpus is about double the size of the ME part of the \textit{HC}. Its great advantage is that it gives complete texts instead of text samples. The user of this corpus should, however, keep in mind that its selection of texts is somewhat narrower than that of \textit{HC}. It does not include poetical texts and some \textit{HC} genres have been omitted: laws, documents and correspondence. These genres would be important in the study of the emergence and establishment of new lexical items and grammatical structures.

A much larger ME corpus, including complete texts, is the seven-million-word \textit{Innsbruck Corpus of Middle English Prose (ICoMEP)}. Even in this corpus the selection of genres is somewhat more restricted than in \textit{HC}: as the name suggests, there are no poetical texts.

By far the largest ME corpus is the \textit{Corpus of Middle English Prose and Verse (CMPEV)} including more than 20 million words. This corpus, part of the \textit{Middle English Compendium} (University of Michigan, Ann Arbor), is excellent in its coverage. Its search program is, however, somewhat awkward: the text files including occurrences of the searched items must be opened one by one.

As to more focused ME corpora, \textit{A Linguistic Atlas of Early Middle English (LAEME)}, compiled at the University of Edinburgh, and the \textit{Middle English Grammar Corpus (MEG-C)}, compiled at the University of Stavanger, consist of Early Middle English (EME) and Late Middle English (LME) texts, respectively.

In the following, the origin and development of two adverbial connectives will be described in order to illustrate the information given by ME corpora: \textit{according to/as} and \textit{concerning}.\footnote{See Rissanen (2000) and Kahlas-Tarkka/Rissanen (forthcoming).} Like most new ME connectives, they are related to verbs borrowed from French.
They represent two different lines of development: according to/as is mainly the result of ME grammaticalization from the verb accord, borrowed in EME, while concerning was borrowed as the grammaticalized connective (OF considerant) in LME.

The verb accord seems to be one of the earliest French loans in ME. In the HC, the earliest instances occur in 12th-century annals of the Anglo-Saxon Chronicle. The same passages can also be found in the Penn-Helsinki Corpus of Middle English (PPCME2) and in the ICoMEP, but not in the much larger CMEPV, which does not include the Peterborough Chronicle. The earliest sense of this loanword refers to an alliance or agreement between two armies or parties (examples (6)-(7)):

(6) & syððan þæs cynges mæn Heanriges manega him to gebugen. & wið hine acordedan (Anglo-Saxon Chronicle annal for 1119, 248 CHRONE2 HC)
‘And afterwards many of King Henry’s men returned to him, and accorded with him’

(7) On þis kinges time wes al unfríð & yfel & ræflac, for agenes him risen sona þa rice men þe wæron swikes, alre fyrst Balduin de Reduers; & held Excestre agenes him & te king it besæt, & siððan Balduin acordede. (Peterborough Chronicle 55 annal for 1135, 55 PETERB HC)
‘In this king’s time was all dissention, and evil, rapine; for against him rose soon the rich men who were traitors; and first of all Baldwin de Redvers, who held Exeter against him. But the king beset it; and afterwards Baldwin accorded.’

The context is military, with reference to alliance. The subject of the verb is personal; the personal/impersonal variation of the subject is of primary importance in tracing the path of grammaticalization of verb-derived connectives. The action indicated by accord in these early examples is rather concrete, although it also includes some cognitive action, decision-making.

The HC gives no other occurrences of the verb accord in 12th- or 13th-century texts. Even larger corpora give rather disappointing results. There is only one 13th-century occurrence in CMPEV, in the Owl and the Nightingale, dating from the second half of the century:
& fo we on mid riʒte dome.
mid faireworde & mid ysome [= ‘peaceful’].
þeʒ we ne bo at one acorde,
we m[a] ʒe bet mid fayre worde,
witute cheste [= ‘strife’], & bute fįʒte,
plaidi [= ‘argue’]/ ... (The Owl and the Nightingale 179-184 CMPEV)

Accord is used as a noun, ‘although we are not in accord’; both the owl and the nightingale, referred to by we, are animate and personified.

The popularization of accord, verb and noun, begins in the 14th century. There are ten instances in four different texts dating from the first half of the 14 century in HC (Table 3).

<table>
<thead>
<tr>
<th>TEXTS</th>
<th>ABSOLUTE NUMBERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early South English Legendary</td>
<td>2</td>
</tr>
<tr>
<td>Robert of Gloucester’s Chronicle</td>
<td>5</td>
</tr>
<tr>
<td>Earliest English Prose Psalter</td>
<td>2</td>
</tr>
<tr>
<td>Ayenbite of Inwyt</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Accord in early 14th century texts in the ME part of the HC. Absolute numbers.

The other ME corpora do not include a substantially larger number of early occurrences. This is not surprising as the ICoMEP and the PPCME2 do not include poetical texts. The scarcity of early 14th-century instances in the CMEPV may be due to the text selection of this corpus which is biased towards Chaucerian and other late 14th- and 15th-century writings.

In some early 14th-century instances the subject is inanimate, as in examples (9)-(10):

(9) “Aue Maria, gracia plena:” þeos four wordes weren i-do
And i-graued [= ‘engraved’] in þis ring: for huy a-cordeden wel þar-to.
(Early South English Legendary 434 SLEG HC)

(10) þou sselt ywyte [= ‘know’] / þet zoʒnesse [= ‘truthfulness’] halt þise riʒtuolle waye. Vor zoʒnesse acordeþ þe onderstondinge of þe herte: (Ayenbite of Inwyt I 256 AYENB HC)
In these examples, *accord* is used to indicate similarity between words and ideas. This development can be explained with reference to the basic meaning of *accord*, which includes the concepts of agreeing and agreement. Similarity of thinking, purpose, activity, etc. is easily expanded to more general aspects of similarity or identity. Similarity between texts or ideas has cognitive implications related to agreement between people. Traugott/Dasher’s (2002) *ITSC* theory, referred to in §1, is attractive in this context. The speaker or writer expands the use of the verb *accord*, from personal agreement and association to more general, figurative aspects of similarity and being alike, not only in opinion but in appearance. The hearer/reader grasps and accepts this generalization and it becomes part of his/her use of language.

In texts dating from the second half of the 14th century, we can also find examples in which *accord* is used impersonally, *it accordeth*. This structure is even further removed from the original use and sense of the verb:

(11) Manie such aduerbis,coniuncctiouns, and preposicionis ben set ofte oon for another, and at fre chois of autours sumtime; and now tho shulen be taken as it *acordith* best to the sentence. (Purvey *Prologue to the Bible* I 60 WYCPROL HC)

(12) “But,” quod sche, “I am certein by many resouns that schrewes ben unsely.”
“*It accordeth,***” quod I. (Chaucer *Boethius* 448 C2 BOETHCH HC)

Examples like this may signal the first steps towards grammaticalization.

Fifteenth century texts stand in evidence of the development resulting in the grammaticalized connective form *according (un)to*. Here the large corpora, *ICoMEP* and *CMEPV* show their strength, as can be seen in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>accord</th>
<th>according (un)to/as</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>HC</em> (ME part)</td>
<td>170</td>
<td>35</td>
</tr>
<tr>
<td><em>ICoMEP</em></td>
<td>c. 1,000</td>
<td>c. 200</td>
</tr>
<tr>
<td><em>CMEPV</em></td>
<td>c. 3,000</td>
<td>c. 400</td>
</tr>
</tbody>
</table>

Table 4. *Accord* and *according to/as* in the ME part of the *HC*, the *ICoMEP* and the *CMEPV*. Absolute numbers.
The most important feature in these 15th-century occurrences is that in many of them the original sense of accord ‘agreeing to’, ‘in agreement with’, is still traceable, although the quality of the subject and the syntactic structure clearly imply that grammaticalization is in progress (examples (13)-(15)):

(13)  In þis tyme, aftir many tretis betwix þis lond and Frauns, and nóþing þat þei profered was according to reson, þe kyng mad redy his schippis at Southampton to spede him to his conquest. (Capgrave 243 CAPCHR HC)

(14)  Justice shuld be had and exercised amonge all yo’ Subgettes w’in this yo’ seid Realme according to theffect of the seid Statutes and orden’nces, (Statutes 551 STAT2 HC)

(15)  frome hensforth the chalenge comenly callid Riens Deyns le garde be no chalenge but utterly voide and of none effecte: Savyng to every person all marner of other challenges what so ever they be according to the lawe. (Statutes 552 STAT2 HC)

In example (13) according can still be regarded as a verb form, ‘nothing that they proffered was according to reason’, but the indication is weakened from ‘agreeing on’ to ‘equivalent to’ and the construction is probably to some extent prepositionalised: the expression sounds idiomatic: “x is according to reason”, “x is reasonable”.

The decisive step from verb to preposition is taken when the link between the subject of the sentence and according is broken. In (14) and (15), the prepositional quality of according (un)to is confirmed by the fact that no clearly definable subject for accord exists. It is a natural result of the shift of focus from the subject of according to the source, authority, etc., expressed by the object or complement of the prepositional according (un)to, as in (15), in which the topic is rather indefinitely expressed by “all manner of other challenges” and specified by the prepositional phrase “according to the law”:

Thus the decisive step in the grammaticalization of according (un)to is the separation of the prepositional phrase from the subject of the verb accord. Schematically, this development can be presented in the following way:
1. x is according to y and is/does z.
2. x according to y is/does z.
3. x is/does z according to y.
4. according to y x is/does z.

The connection between the subject represented by x and the verb form *according* is weakened and the focus is shifted to the object or complement y.

By the end of the 15th century, *according (un)to* can be regarded as a grammaticalized preposition. A survey of the genres of texts in which this new prepositional use appears is interesting (Table 5).

<table>
<thead>
<tr>
<th>ME4 (1420-1500): GENRES</th>
<th>ABSOLUTE NUMBERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statutes</td>
<td>13</td>
</tr>
<tr>
<td>Petitions</td>
<td>1</td>
</tr>
<tr>
<td>Handbooks</td>
<td>4</td>
</tr>
<tr>
<td>Sermons</td>
<td>5</td>
</tr>
<tr>
<td>Religious treatises</td>
<td>6</td>
</tr>
<tr>
<td>Prefaces (Caxton)</td>
<td>1</td>
</tr>
<tr>
<td>History</td>
<td>1</td>
</tr>
<tr>
<td>Fiction</td>
<td>2</td>
</tr>
<tr>
<td>Mystery</td>
<td>1</td>
</tr>
<tr>
<td>Priv. correspondence</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>35</strong></td>
</tr>
</tbody>
</table>

Table 5. *According (un)to/as* in the sub-period ME4 (1420-1500) of the *HC*. Absolute numbers.

The figures show that more than one third of the occurrences can be found in statutory texts, although these texts only represent some ten per cent of all the texts of this sub-period. Generally speaking, official language played an important role in the development of new grammaticalized connectives. This is not surprising as accurate ways of indicating relationships between clauses and propositions was of great importance in official writing. For this reason, the inclusion of texts representing officialese could be recommended in all corpora giving information on the development of English in the vitally important late 14th and 15th centuries.
One feature worth noting in the development of grammatical vocabulary is the proportional frequency of the grammaticalized use in comparison with the other, non-grammaticalized uses of the same word. As can be seen in Table 6, in the 15th century the frequency of the uses of the verb *accord* was much higher than that of the phrase *according (un)to*. In EModE, in the 16th century, the proportion changes radically.

<table>
<thead>
<tr>
<th>ME4 (1420-1500)</th>
<th>EModE1 (1500-1570)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>According (un)to/as</em></td>
<td>35</td>
</tr>
<tr>
<td><em>Accord (other forms)</em></td>
<td>57</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>92</strong></td>
</tr>
</tbody>
</table>

Table 6. *Accord* and *according (un)to/as* in the sub-periods HM4 (1420-1500) and EModE1 (1500-1570) of the HC. Absolute numbers.

While the majority of the 92 ME4 occurrences represent non-grammaticalized uses (57 as against 35), no less than 67 out of the 79 occurrences in EModE1 are grammaticalized phrases *according (un)to* or *according as*. The grammaticalized or rapidly grammaticalizing forms have become clearly more common than the verbal uses. The *Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)* (Kroch et al. 2004) includes statutes and private correspondence, and its genre selection is thus comparable with that of the *HC*. The usefulness of this 1.8 million-word corpus is obvious in confirming the results of the much smaller EModE part of *HC*. There are no less than 233 instances of *accord(ing)*, and 194 of those are of the grammaticalized forms.

It seems that the grammaticalization of an item can be regarded as completed when the grammaticalized form or forms clearly outnumber the other uses of the same base word. Judging by this criterion, corpus evidence suggests that the prepositional connective *according (un)to* was established in the 16th century. It is possible that the French expression *accordant que* supported this development, but in all probability the grammaticalization essentially took place in ME.

---

7 The subordinator *according as* appears in the 15th century, but it seems to be established only in EModE. There are only three instances in HM4 and nine in HE1.
This is suggested by the very frequent use of the verb *accord* before the establishment of the connective and the dramatic drop of the verbal uses in the Early Modern period.

4. Early Modern English

As mentioned above, a different line of development can be traced in the history of another verb-based adverbial connective, the preposition *concerning*. The verb *concern*, borrowed from French, only appears in English at the very end of the ME period and it seems that the connective *concerning* was borrowed at the same time as the verb, or even earlier. The earliest instances are recorded in texts dating from the second half of the 15th century (examples (16)-(18)):

(16) And yif eny suyt wer commenced afer that tyme hadd for cause of eny such seisyng in which eny issue (*concernyng* þat acte sh)uld be taken þat the seid issue shuld be tried in the shire (*Petitions* 269 PET4 HC)

(17) And I bad hym say a-gayn þat I com thyder to hym fore hijs own worchype and auayle, and þat I wasse sory þat I com so fere fore hym. And after þat he sent fore me and he cowde not fynde me, and I harde say there-of; and þan I wrott a letter resytyng how þat he wasse sworn ȝesterday fore to say þe trowthe of al maner of materis *concernyng* Sire John Fastolfe, auysyd hym to remembere qwat hijs wytnesse hadde sayd fore hijs sake (*Paston Letters* 203 CPASTON HC)

(18) buylved by Salomon. ne in buyldyng ne yet in apparell as Josephus declareth / For in yȝe fyrst temple was conteyned the arke of the olde testament. in it *concernyng* the tables of the lawe. Manna & the rodde of Aaron. whiche flourisshid & fructified by myracle. (*Fitzjames* A5V SERM HC)

There are no instances of the other uses of *concern* in the ME part of the *HC*. The numbers of occurrence in Table 7 suggest that the connective appeared first, borrowed from French as a grammatical word and that it supported the introduction and establishment of the non-grammaticalized uses of the borrowed verb.
Concerning (connective) 4 31 56
Concern (verbal uses) 0 15 12

Table 7. Concern(ing) in the Middle English part of the HC, the ICoMEP and the CMEPV. Absolute numbers.

ICoMEP and CMEPV support the impression given by the small number of instances in HC: from the very beginning, the connective concerning is more common than the verbal uses of concern. All instances come from 15th-century texts.

The EModE part of HC shows the increasing popularity of the preposition compared to the occurrences of the other uses of concern (Table 8).

<table>
<thead>
<tr>
<th>Concerning (connective)</th>
<th>1500-1570</th>
<th>1570-1640</th>
<th>1640-1700</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Handbook</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Science</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Education</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Philosophy</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Sermon</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Trials</td>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>History</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Travelogue</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Diary</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Biography</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Fiction</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Comedy</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Correspondence - private</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Correspondence - official</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Bible</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>32 (16.8)</td>
<td>34 (17.9)</td>
<td>43 (25.1)</td>
</tr>
</tbody>
</table>

Table 8. Concern(ing) in the EModE part of the HC. Absolute numbers. (Numbers per 100,000 words in brackets).

The genre distribution shows that the connective is spread to various text types, including fiction and comedy (examples (19) and (20)).
(19) To whom a yonge man came to speke wyth her in her ere as though he hyt had bene for som matre concerning the funerallys/ howe be yt he spake of no such matter but only wowyd her that he myghte be her husbande/ (Merry Tales 21 MERTYATAL HC)

(20) To be taken for a woman enclined to vice. Nothing at all to Vertue gyuing hir due price, Wherfore concerning mariadge, ye are thought Suche a fine Paragon, as nere honest man bought. (Roister Doister 1077 UDALL HC)

The bias is, however, still on the formal side: the twelve instances in the trial records and the five in official correspondence are worth noting. In the 17th century, the verbal uses increase quite remarkably: out of the 71 occurrences in texts dating between 1640 and 1710, 28, or about 40 per cent, represent verbal uses. The proportional increase of the use of the grammaticalized concerning is also worth noting: from 17 to 25 per 100,000 words.

There are highly useful EModE corpora focusing on more narrowly defined genres, which give additional information on the development and establishment of adverbial connectives: among others, the Parsed Corpus of Early English Correspondence (PCEEC), the Corpus of Early English Medical Writing (CEEM), and the Corpus of English Dialogues (CED). CEEC and CEEM cover the time from LME to LModE, although the latest periods are still under construction.

The PCEEC is particularly important as it shows how items and constructions first appearing in more formal and literary genres spread to private letters, often more rapidly than could be expected (Table 9).

<table>
<thead>
<tr>
<th>ITEMS</th>
<th>ABSOLUTE NUMBERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerning (connective)</td>
<td>29</td>
</tr>
<tr>
<td>Concern (verbal use)</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 9. Concern(ing) in the PCEEC (early letters, c. 1400-c. 1520). Absolute numbers.

It is also important to note that the total of 29 occurrences of the connective in 15th and very early 16th century letters represents roughly ten occurrences per 100,000 words. One explanation to the relatively
high frequency is no doubt the fact that these letters often discuss rather formal aspects, such as legal matters, wills, etc. But private letters also form an important link in introducing changes from above to the more common and varied areas of human communication. Here, once again, the concept of \textit{IITSC} may be helpful in understanding the path of the connective from official contexts via private letters to more general use.

The \textit{CED} and the \textit{CEEM}, represent in a way, two opposite ends of discourse (see Tables 10 and 11):

<table>
<thead>
<tr>
<th></th>
<th>\textit{CED1} (1560-1599)</th>
<th>\textit{CED2} (1600-1639)</th>
<th>\textit{CED3} (1640-1679)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{concerning}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trials</td>
<td>6</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>Witness depositions</td>
<td>11</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>Handbooks</td>
<td>9</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Comedy</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Fiction</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Miscelanea</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>28 (14.0)</td>
<td>22 (10.6)</td>
<td>60 (23.1)</td>
</tr>
<tr>
<td>\textit{concern (other uses)}</td>
<td>5</td>
<td>9</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 10. \textit{Concern(ing)} in the \textit{CED}. Absolute numbers.

As can be seen in Table 10, the \textit{CED} consists of a number of genres including dialogue. Even in this corpus, there is a notable increase in both verbal and prepositional uses in the 17th century. The somewhat surprisingly high rate of occurrence in dialogue texts is due to the high frequencies in Law Court Trials and Witness depositions. Even in focused corpora there may be different genres or types of text.

<table>
<thead>
<tr>
<th></th>
<th>15th c.</th>
<th>1500-1599</th>
<th>1600-1699</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Concerning}</td>
<td>1</td>
<td>40</td>
<td>348</td>
</tr>
<tr>
<td>\textit{Concern (other uses)}</td>
<td>0</td>
<td>2</td>
<td>101</td>
</tr>
</tbody>
</table>

Table 11. \textit{Concern(ing)} in the \textit{CEEM}. Absolute numbers.

The \textit{CEEM} opens a window to LME and EModE scientific writing. It is not surprising that there are hardly any occurrences in the 15th century, represented by \textit{Middle English Medical Texts (MEMT)}; in 16th-
century texts, practically all occurrences of *concern* are of the grammaticalized form, and the increase in the 17th century of both prepositional and verbal uses is remarkable.

As to corpora giving information on the regional varieties of English, the general pattern of the distribution of the occurrences in the 800,000-word *Corpus of Older Scots* corresponds to that of the *HC*. There are very few pre-1500 instances, and in the Early Modern period the preposition is clearly more popular than the verbal uses. The structure of the Scots corpus is roughly similar to that of *HC*.

5. Late Modern English

The occurrences of *concerning* in LModE, from the 18th century on, can be studied using very large – even gigantic – corpora. The most important ones are the *Corpus of Late Modern English Texts, Extended Version (CLMETEV)*, *A Representative Corpus of Historical English Registers (ARCHER)*, the *Penn Parsed Corpus of Modern British English (PPCMBE)*, and the *Corpus of Historical American English (COHA)*.

*CLMETEV* covers the time from the early 18th century to the 1920s. Its size is around 15 million words, and it is divided into three subcorpora on chronological basis. The *CLMETEV* occurrences suggest that the overall popularity of *concern* decreases in the course of the 18th and 19th centuries (Table 12). The prepositional use becomes more infrequent than the other uses of *concern*. This seems to indicate the narrowing of the genres in which *concerning* is favoured: stylistic or register-based restrictions.

<table>
<thead>
<tr>
<th></th>
<th>1710-1780</th>
<th>1780-1850</th>
<th>1850-1920</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerning</td>
<td>545 (17.9)</td>
<td>452 (7.9)</td>
<td>333 (5.3)</td>
</tr>
<tr>
<td>Concern (other uses)</td>
<td>636 (20.9)</td>
<td>665 (11.6)</td>
<td>602 (9.6)</td>
</tr>
</tbody>
</table>

Table 12. *Concern*(*ing*) in the *CLMETEV*. Absolute numbers. (Numbers per 100,000 words in brackets).
ARCHER is much smaller than CLMETEV, only 1.8 million words, but it is structured by genres and consists of both British and American English (BrE, AmE) texts. The numbers confirm the decreasing trend from the 18th to the 20th century; the proportional increase of the verbal forms and the decrease of the connective is roughly similar to that indicated by CLMETEV (Table 13).

<table>
<thead>
<tr>
<th></th>
<th>1700-1799</th>
<th>1800-1899</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerning</td>
<td>61 (11.4)</td>
<td>28 (5.2)</td>
</tr>
<tr>
<td>Concern (other uses)</td>
<td>97 (18.1)</td>
<td>60 (11.1)</td>
</tr>
</tbody>
</table>

Table 13. Concern(ing) in ARCHER. Absolute numbers. (Numbers per 100,000 words in brackets).

The million-word PPCMBE covers the 18th and 19th centuries; the last text dates from 1914. There are 122 occurrences of the form concerning and 110 of the other uses of concern. The gradual proportional increase of the verbal uses can be seen in these numbers.

The size of the COHA is most impressive: 400 million words covering the period from 1810 to 2009 (Table 14). There is a rough genre division in this corpus.

<table>
<thead>
<tr>
<th></th>
<th>1810-1900</th>
<th>1910-2000</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerning</td>
<td>8,164 (5.3)</td>
<td>7,523 (2.8)</td>
<td>15,687 (3.9)</td>
</tr>
<tr>
<td>Concern (other forms)</td>
<td>11,179 (7.5)</td>
<td>46,263 (18.5)</td>
<td>57,442 (14.4)</td>
</tr>
</tbody>
</table>

Table 14. Concern(ing) in the COHA. Absolute numbers. (Numbers per 100,000 words in brackets).

Table 14 shows a rather similar development as that indicated by the other LModE corpora. The high frequency of the verbal uses in 20th-century texts is worth noting. There may be a difference in usage between BrE and AmE texts worth further study.
6. 20th century

The Brown corpus family of late 20th-century BrE and AmE (Brown, Frown, LOB, F-LOB) gives corresponding results (Table 15). These four corpora have one million words each and the numbers are thus easy to compare. The low frequency of concerning, compared with the other uses of concern is obvious.

<table>
<thead>
<tr>
<th></th>
<th>Brown (AmE 1960s)</th>
<th>Frown (AmE 1990s)</th>
<th>LOB (BrE 1960s)</th>
<th>F-LOB (BrE 1990s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerning</td>
<td>62</td>
<td>47</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>Concern (other forms)</td>
<td>276</td>
<td>279</td>
<td>303</td>
<td>328</td>
</tr>
</tbody>
</table>

Table 15. Concern(ing) in late 20th century BrE and AmE Corpora. (Absolute numbers).

Finally, the occurrences in two very large PDE corpora: the British National Corpus (BNC, 100 million words) and the Corpus of Contemporary American English (COCA, over 450 million words) can be seen in Table 16.

<table>
<thead>
<tr>
<th></th>
<th>BNC (AmE 1960s)</th>
<th>COCA (AmE 1990s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerning</td>
<td>3,279 (3.3)</td>
<td>11,370 (2.52)</td>
</tr>
<tr>
<td>Concern (other forms)</td>
<td>30,092 (30.1)</td>
<td>105,335 (23.4)</td>
</tr>
</tbody>
</table>

Table 16. Concern(ing) in the BNC and the COCA. Absolute numbers. (Numbers per 100,000 words in brackets).

The overall picture given by the numbers in Table 16 corresponds roughly to the one given by the Brown family corpora in Table 15. The somewhat lower frequencies in COCA, when compared with Brown and Frown is, however, worth noting.
7. Conclusions

In this paper the long-diachrony development of two adverbial connectives, according (un)to and concerning have been described on the basis of the information given by historical corpora. Although both items are based on French loanwords, their appearance in ME follows two different paths: according (un)to is mainly a result of the grammaticalization of the -ing form of the verb borrowed in very early ME, while concerning was borrowed as a preposition; the other forms of the verb concern emerge in LME more or less at the same time as the connective. The ModE development of concerning shows how the verbal forms gradually supersede the connective in frequency and how concerning loses ground in LModE and becomes stylistically marked. This development probably explains the incomplete grammaticalization of concerning: some dictionaries mark concerning as “formal” and Quirk et al. (1985: 660) label it as “marginal preposition”.

The role of general and focused corpora in the variation-based long-diachrony study of the language is obvious. Study of changing frequencies in the course of centuries is now much easier than it was before the time of corpora, and the influence of genre-based, regional and sociolinguistic external factors can be more reliably analysed. It is obvious that focused small corpora, underlining change through variation, bring us closer to the actual spoken and written expression of the past centuries. Large general corpora draw the overall picture of development and change over the centuries and provide the scholar with more substantial quantitative results. Finally, scholars and students should remember that even the best corpus can never give a fully reliable reproduction of linguistic reality – even at best, it offers only a slice of this reality. Without the slightest doubt, however, corpora have opened us new vistas for understanding and interpreting language history.
On English historical corpora and the development of adverbial connectives

References

Some important English historical corpora

1. Old English

*DOEC (Dictionary of Old English Corpus)* (c. 3.5 million words) = diPaolo Healey et al. (2009) and diPaolo Healey (ongoing).


*YCOE (York-Toronto-Helsinki Parsed Corpus of Old English Prose)* (c. 1.5 million words) = Taylor et al. (2003).

*SCONE (Seville Corpus of Northern English)* (7th-16th cc.) = Fernández Cuesta et al. (2008-).

2. Middle English


*PPCME2 (Penn-Helsinki Parsed Corpus of Middle English 2)* (c. 1.1 million words) = Kroch/Taylor (2000).

*ICoMEP (Innsbruck Corpus of Middle English Prose)* (c. 7 million words) = Markus (2010).

*CMEPV (Corpus of Middle English Prose and Verse)* (c. 20 million words) = Humanities Text Initiative, University of Michigan (2006).

*LAEME (Linguistic Atlas of Early Middle English)* (c. 650,000 words) = Laing (2008).

*MEG-C (Middle English Grammar Corpus)* (Stavanger) (c. 450,000 words) = Stenroos et al. (2008-)

3. Early Modern English


*PPCEME (Penn-Helsinki Parsed Corpus of Early Modern English)* (1.8 million words) = Kroch et al. (2004).

*PCEEC (Parsed Corpus of Early English Correspondence)* (c. 2.2 million words) = Taylor et al. (2006).
CED (Corpus of English Dialogues) (1.2 million words) = Kytö/Culpeper (2006).
EMEMT (Corpus of Early Modern English Medical Texts) (c. 2 million words) = Taavitsainen/Pahta (2010).
HCOS (Helsinki Corpus of Older Scots) (c. 800,000 words) = Meurman-Solin (1995).

4. Late Modern English

CLMETEV (The Corpus of Late Modern English Texts, Extended Version) (c. 15 million words) = De Smet (2006).
ARCHER (A Representative Corpus of Historical English Registers) (c. 1.7 million words) = ARCHER.
PPCMBE (The PennParsed Corpus of Modern British English) (c. 1 million words) = Kroch et al. (2010).
COHA (The Corpus of Historical American English) (c. 400 million words) = Davies (2010-).

5. Present-day English

F-LOB (The Freiburg-Lancaster-Oslo/Bergen Corpus) (1990s; 1 million words) = Mair/Leech (2007).
BNC (The British National Corpus) (100 million words) = BNC.
COCA (The Corpus of Contemporary American English) (450 million words) = Davies (2008-).
Studies

ARChER = A Representative Corpus of Historical English Registers. 1993-. <manchester.ac.uk/archer>, <http://www.helsinki.fi/varien g/CoRD/corpora/ARChER/index.html>.


Measuring typological syntheticity of English diachronically with the use of corpora

Abstract

The paper proposes a quantitative, but holistic methodology for establishing a level of morphological syntheticity. The methodology is based on a series of corpus-based probes into morphological behaviour of selected high-frequency nouns, adjectives and verbs from Old English to Present-day English. Reflecting on the distribution of the morphological markers in the four periods under scrutiny, levels of syntheticity are established that correspond to the textbook story of a typological re-shaping in the history of English.

1. Preliminaries

The present paper has been inspired by Benedikt Szmrecsanyi’s (forthcoming) corpus-based study investigating the varying levels of overt grammatical analyticity as opposed to overt grammatical syntheticity in the coding of grammatical information from Early Middle English (EME) to Present-day English (PDE). In an attempt to test the usability of historical corpora as methodological tools in this area of research, we propose to complement and refine Szmrecsanyi’s methodology and criteria of corpus-based morphological analysis in the framework of quantitative morphological typology.
2. Szmrecsanyi’s analysis

Szmrecsanyi’s probe offered a macroperspective analysis, based on holistic measurements of the analyticity-syntheticity dichotomy in English between 1100 and 1900. These were carried out in terms essentially established by Greenberg (1960), with grammatical analyticity defined as the text frequency of free grammatical markers, and grammatical syntheticity defined as the text frequency of bound grammatical markers. Szmrecsanyi’s “analyticity index” had been calculated as the ratio of the number of free grammatical markers (i.e., function words) in a text to the total number of words in the text, normalized to a sample size of 1,000 tokens, whereas his “syntheticity index” had been defined as the ratio of the number of words in a text that bear a bound grammatical marker to the total number of words in the text, again normalized to a sample size of 1,000 tokens.

Tapping the Penn Parsed Corpora of Historical English series, Szmrecsanyi’s analysis was concerned with token frequencies, rather than, e.g., the size of inventory types, in order “to avoid intuitive estimates based on overall impressions” (Greenberg 1960: 185, quoted by Szmrecsanyi, forthcoming: 657). His calculations had brought a surprising result, one that essentially undermines the received story of the typological re-shaping of English from synthesis to analysis: Szmrecsanyi found out that analyticity was on the rise all the way from the beginnings of Middle English (ME) until the end of the Early Modern English (EModE) period, but declined subsequently; the reverse he showed to be true of syntheticity. Szmrecsanyi’s novel methodological approach had thus brought surprising results, in contrast to the traditional typological view of the history of English, which sees it as an exemplary instance of a progression from a predominantly synthetic (Old English, OE) to a predominantly analytic (PDE) type of language (cf. e.g., Skalička/Sgall 1994 or Kastovsky 2006).

---

2 See <http://www.ling.upenn.edu/hist-corpora>. 
3. Aims and Sources

Since there is no generally established measure of syntheticity and since we focus on measuring the syntheticity of diachronic varieties of a single language, we have decided that we need a more refined method of measurement than Szmrecsanyi’s, whose focus seems to be more general: his analysis uses the same method to compare different languages as well as the different varieties and dialects of the same language. At the same time we are aware that our own method would benefit from a further application to a wider selection of varieties or languages. Without such comparison, it is difficult to assert the overall significance of relative distances measured between the varieties under scrutiny.

The present paper aims to complement Szmrecsanyi’s research as well as methodology by a series of corpus-based probes into morphological behaviour of selected high-frequency nouns, adjectives and verbs. We have opted to measure syntheticity only, but have extended the time scope by the OE period, thus tracing the levels of syntheticity in English through all the major historical periods. As regards our choice of corpora, the OE period has been covered with the help of the York-Toronto-Helsinki Parsed Corpus of Old English Poetry (Taylor et al. 2001) and the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE; Taylor et al. 2003), and Present-day English (PDE) has been analysed using the British National Corpus (BNC) – as an alternative source to The Penn Parsed Corpus of Modern British English (PPCMBE) (Kroch et al. 2010, used by Szmrecsanyi forthcoming).

4. Methodology

These corpus-based probes have helped us to see how syntheticity has been mapped and maintained between form and function across time.
Rather than focusing on frequencies of individual exponents (for example, that of the genitive singular) separately, we have looked into the distribution of markers across paradigms as principal organizational devices. Presuming that the more varied the inflectional system of a language is, the more syntheticity the language manifests, we have proposed to measure such variability by sampling the frequencies of select word forms classified into formal bundles of categories and by computing a syntheticity index – first for each sampled paradigm, and then across all the sampled data.

Our analysis was centred on calculating the distribution of morphological markers of grammatical inflection (syntheticity) based on corpus-derived word forms/tokens by classing them into paradigmatic categories defined primarily according to form. The use of tagged corpora enabled us to base our analysis consistently on word-class membership, but the inclusion of class/paradigm membership into our methodology meant that we had to retrieve all variants (inflectional or otherwise) of each word. With non-lemmatized corpora for the three earlier periods the only feasible method was manual processing of all the data (both in retrieving it and in classifying it). For this reason we had to limit our probe to a representative selection of words. Because we were not aware of any reliable data on word-frequencies for the earlier periods (again probably due to non-existence of lemmatized and standardized corpora), the selection was based on a) traditional grammars of OE (which usually provide some inkling of the overall frequency of each class and also list mostly the more frequent members of such classes); b) a preliminary probe in which we discarded words that were not steadily above 0.1 per million across the periods.

4.1. Selecting the representative list of words

The list of the words was based on the following criteria:

- high frequency of items in all the periods under scrutiny;
- items of the same word-class and functionally stable (including inflected forms);
• low degree of homonymy with other items of the same word-class; and
• representativeness in the given inflectional subsystem of the OE period (such as the choice of \(a\)-stems, \(n\)-stems and \(ō\)-stems, rather than \(i\)-stems or consonant stems, ultimately based on their rough quantitative share in the nominal system).³

4.2. Capturing the word-forms

Using data about attestation from the Dictionary of Old English (DOE) (Healey 2008), An Anglo-Saxon Dictionary Online (Bosworth et al. ongoing), the Middle English Dictionary (MED) (Kurath et al. 1952-2001) and the Oxford English Dictionary Online (Simpson 2000-), and combining them with rules of inflectional morphology as described in standard grammars of the respective periods,⁴ regular expressions⁵ were constructed that captured all possible members of inflectional paradigms (including spelling or dialectal variation) for each of the selected words in the respective corpus. A total of over 913,098 tokens was considered: between 26,326 tokens for OE; 22,852 for ME; 29,107 for EModE; and 834,813 for PDE.

4.3. Categorization

The paradigmatic categories were outlined separately for each word-class. Their definition, as well as the membership of word forms in the categories, came to be based on a combination of formal and functional criteria, although the formal criteria were primary. The following principles were observed when classifying the data:

---

³ E.g., Wright (1925) or Campbell (1983).
⁴ Such as the grammars of OE by Wright (1925) and Campbell (1983).
⁵ Such as “[eiy][oa]?:[ťd̥t][h]?:...?” to capture forms of the word earth in the OE periods. Sometimes, more than one query was necessary to capture all possible candidate forms and unwanted results had to be discarded manually.
Tokens were classed according to their formal characteristics disregarding their function (e.g., with the nominal -an in OE no attempt was made to distinguish between, say, the functions of accusative singular and nominative plural).

Therefore, if any of the traditional functional categories agreed in their formal characteristics across the whole word-class, they were merged together disregarding their functional differences (e.g., OE -V-e for 2nd singular preterite indicative and subjunctive).

If a token differed from the standard formal description of a paradigmatic category, or was found to be formally ambiguous in a manner testifying to a probable dialectal or spelling variation, but at the same time agreed with the functional description of the slot in question and, in addition, was not formally homonymous with a different category (retaining, therefore, a formal basis for the functional distinction), it was classed under the given slot (e.g., OE hulpan had been classed as plural preterite indicative, because the -an/-on variation was common, especially in Late Old English (LOE), and the token was formally distinguished from the forms of the -an category by its root vowel).

If a number of formally similar tokens across different functional categories showed little regular formal differentiation with regard to their functions, the formal differentiation of the categories was considered untenable and the categories had been merged (e.g., -ra or -re in OE adjectival comparison and inflection; or EModE and PDE -0 or -e for nominative and dative).

As a result, eight paradigmatic categories were established for nouns, 16 for adjectives and 22 for verbs (see Tables 2-4 for the formal characteristics of the resulting categories). As follows from the principles above, our definition of the categories was based on a combination of formal and functional criteria. Moreover, because the definition had been designed to span all the periods, it was bound to be to a degree arbitrary. To exemplify this: on the one hand, the only category not spanning all the periods is the final -e, which we considered functionally empty in the two latter periods and had, therefore,
merged it with final -0 there. On the other hand, the verbal categories expressed by the final -de and -te we considered to be sufficiently distinct, especially in the later periods, for them to be kept separate.

5. Problems in generating the data

Even in collecting a relatively limited sample of data with the help of historical corpora, analysis is bound to face a range of problems. Some are of a more general and obvious nature, in that they inevitably beset corpus-based computing of this kind, such as uneven frequency levels of items required (with particularly low numbers in the early segments of history) and sifting through spelling and dialect variation in ME (with the specific difficulty of the spelling -e, which interferes with the sound of the day and which had to be taken, for the purposes of our analysis, at its face value).

On the other hand, some methodological difficulties encountered in producing our sample were more specific, linked to individual lexical items. Of such nature was, for instance, disambiguation of homographs (e.g., distinguishing between such items as hēah and hēan in OE or between full and foul in ME, which had to be done manually) and accommodating period frequency swells. These were, in most cases, due to varying distribution of the item under scrutiny within its lexical field in the given period (such as the decreasing occurrence of the adjective micel – as opposed to grete – or sellan meaning ‘to give’ in the ME period), or, in contrast, due to an increase in the frequency of an item which was caused by a very high profile of prepositional phrases based on it (such as the high frequency of the prepositional phrase withouten ende in ME).
6. Measuring syntheticity

One of the major questions in our research was how to measure syntheticity as based on the frequency of the morphological markers. Our understanding of syntheticity stems from both the frequency of synthetic markers and their distribution throughout the inflectional system: we consider a language with two very frequent inflectional markers intuitively less synthetic than a system with four inflectional markers of the same overall frequency. In fact, the greater the choice a speaker of a language has, when choosing an inflectional marker for a random word, the greater the syntheticity of the particular language. This wider selection scope is then manifested as a higher frequency of the choices in the corpus.

To capture this freedom of choice, or how well distributed the markers are throughout the system, we employed the informational (or Shannon) entropy, which “[…] is a measure of the degree of randomness (or ‘shuffledness’ if we think of a pack of cards)” (Oakes 1998: 58). The formula chosen was the standard Shannon entropy formula, where \( n \) is the number of formal categories (as defined above), \( x \) is the number of realizations of a given category in a given period and \( p \) is the probability of such realization (there were, for example, 340 realizations of the “third person singular present -eth/es” for verbs in the OE period, which gives it a 6.15% probability of realization within the total 5,526 verb forms examined for the period):

\[
- \sum_{i=1}^{n} p(x_i) \log(p(x_i))
\]

The resulting number is an index of syntheticity, where the higher value indicates a higher level of syntheticity. It is not very meaningful on its own, but rather more useful if considered as a relative value for comparison of different language varieties.

To calculate the index, the number of tokens for each category and period of each word was taken into account (see Table 1) and the total number of tokens collected for the whole word-class, period and
category was divided by the total number of tokens for the word-class in that given period, thus producing the percentage distribution for a given category and period in the whole word-class. From the percentages, an overall index of syntheticity for a given word-class and period was calculated (see Tables 2-4). This way, if a particular word generally occurs more frequently in any period, it is found to contribute more significantly to the index. It might be more representative to offset the weight of individual word frequency by overall frequency of each paradigmatic category here, but such data are, to the best of our knowledge, not available. Some methodological questions remain open as to the best way the overall index for each world-class should be aggregated. The index may better capture the real syntheticity levels of the whole language system, if we also allowed for the general probability of an individual word to appear in the language, rather than just for the probability of a marker. However, in the case of a relatively limited list of words like ours, this would cause a disproportionate contribution to the index by highly frequent words and could potentially confuse the overall results, because the relative distribution of the words changes between periods (e.g., forms of the word high constitute some 5% of OE sample of adjectival forms, but almost 20% of the PDE sample). In other words, the question is whether the words we selected are representative in their frequency of paradigmatically similar words in the whole language system. With this selection, we are inclined to believe that they are not and that a much larger sample of words would be needed. However, from some preliminary calculations in this direction, we assume that this does not have an important impact on our results.

<table>
<thead>
<tr>
<th>WINTER</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0</td>
<td>174</td>
<td>11</td>
<td>117</td>
<td>7066</td>
</tr>
<tr>
<td>-s</td>
<td>31</td>
<td>0</td>
<td>7</td>
<td>306</td>
</tr>
<tr>
<td>-e</td>
<td>17</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-an</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-a</td>
<td>546</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-um</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-ena</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-u</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Token distribution of winter.
<table>
<thead>
<tr>
<th>Nouns</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0</td>
<td>40.84%</td>
<td>45.05%</td>
<td>70.75%</td>
<td>63.28%</td>
</tr>
<tr>
<td>-s</td>
<td>9.34%</td>
<td>18.13%</td>
<td>28.79%</td>
<td>36.67%</td>
</tr>
<tr>
<td>-e</td>
<td>19.43%</td>
<td>36.37%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-an</td>
<td>15.4%</td>
<td>0.34%</td>
<td>0.46%</td>
<td>0.05%</td>
</tr>
<tr>
<td>-a</td>
<td>9.34%</td>
<td>0.11%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-um</td>
<td>5.47%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ena</td>
<td>0.17%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-u</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Syntheticity</td>
<td>0.69</td>
<td>0.46</td>
<td>0.27</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Table 2. Distribution of nominal markers.

<table>
<thead>
<tr>
<th>Verbs</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0</td>
<td>8.52%</td>
<td>8.35%</td>
<td>47.44%</td>
<td>44.77%</td>
</tr>
<tr>
<td>-te</td>
<td>4.54%</td>
<td>23.12%</td>
<td>27.33%</td>
<td>14.48%</td>
</tr>
<tr>
<td>-e</td>
<td>32.37%</td>
<td>31.24%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-(e)th/s</td>
<td>6.15%</td>
<td>7.30%</td>
<td>5.58%</td>
<td>6.13%</td>
</tr>
<tr>
<td>-ed</td>
<td>13.12%</td>
<td>1.90%</td>
<td>2.31%</td>
<td>6.94%</td>
</tr>
<tr>
<td>-en</td>
<td>0.76%</td>
<td>12.36%</td>
<td>4.09%</td>
<td>4.47%</td>
</tr>
<tr>
<td>-ing</td>
<td>0.00%</td>
<td>1.39%</td>
<td>4.98%</td>
<td>13.56%</td>
</tr>
<tr>
<td>-V3-o</td>
<td>4.69%</td>
<td>3.19%</td>
<td>2.32%</td>
<td>5.36%</td>
</tr>
<tr>
<td>-V4-e</td>
<td>0.47%</td>
<td>3.96%</td>
<td>5.29%</td>
<td>4.28%</td>
</tr>
<tr>
<td>-an</td>
<td>10.79%</td>
<td>0.08%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ath</td>
<td>5.23%</td>
<td>0.05%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-don</td>
<td>5.03%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-V2-en</td>
<td>0.24%</td>
<td>3.06%</td>
<td>0.16%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ton</td>
<td>2.32%</td>
<td>0.82%</td>
<td>0.07%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-V4-on</td>
<td>1.85%</td>
<td>0.92%</td>
<td>0.18%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ande</td>
<td>1.34%</td>
<td>0.77%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-st</td>
<td>1.01%</td>
<td>0.87%</td>
<td>0.15%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-epest</td>
<td>0.87%</td>
<td>0.03%</td>
<td>0.01%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-anne</td>
<td>0.40%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-V4-en</td>
<td>0.05%</td>
<td>0.26%</td>
<td>0.06%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-den</td>
<td>0.13%</td>
<td>0.18%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-test</td>
<td>0.13%</td>
<td>0.15%</td>
<td>0.02%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Syntheticity</td>
<td>0.98</td>
<td>0.89</td>
<td>0.66</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 3. Distribution of verbal markers
Measuring typological syntheticity of English diachronically with corpora

<table>
<thead>
<tr>
<th>ADJECTIVES</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0</td>
<td>15.75%</td>
<td>46.61%</td>
<td>86.98%</td>
<td>79.73%</td>
</tr>
<tr>
<td>-e</td>
<td>14.93%</td>
<td>46.31%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ra</td>
<td>6.36%</td>
<td>3.55%</td>
<td>7.55%</td>
<td>15.90%</td>
</tr>
<tr>
<td>-an</td>
<td>32.62%</td>
<td>0.06%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-est(a)</td>
<td>1.75%</td>
<td>2.35%</td>
<td>5.46%</td>
<td>4.37%</td>
</tr>
<tr>
<td>-a</td>
<td>10.97%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-um</td>
<td>9.49%</td>
<td>0.94%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ne</td>
<td>3.94%</td>
<td>0.03%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-st-an</td>
<td>1.45%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-es</td>
<td>1.34%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-r-an</td>
<td>0.91%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ena</td>
<td>0.23%</td>
<td>0.06%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-u</td>
<td>0.11%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-st-um</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-r-um</td>
<td>0.08%</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SYNTHETICITY</td>
<td>0.86</td>
<td>0.43</td>
<td>0.21</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Table 4. Distribution of adjectival markers.

The overall syntheticity for all word-classes was calculated by adding up the syntheticity index values for each word-class and period and multiplying them by the proportion by which each of the word-classes was represented in the corpus of a given period. If the initial selection of the words is taken as representative of the inflectional systems in all the periods, the trend in the change of the level of syntheticity is now easy to represent. In Figure 1, the levels of syntheticity for individual word-classes and the total levels are represented by different markers, while the linear trend is indicated by the thin black line.

To verify the statistical significance of the trends displayed in Figure 1, we have also compiled an overall table (see Table 5) with a syntheticity index for each word and period examined (here offset by the frequency of the word in a given period, rather than the whole word-class). Based on this table, an ANOVA test has been carried out showing a high statistical significance of the trend, with \( p = 1.13\times10^{-23} \). We have also carried a t-test on each of the two consecutive periods (OE-ME, ME-EModE, EModE-PDE) with the result indicating that

---

6 Based on the POS tagging of the respective corpora.
the differences between the consecutive periods are statistically significant, except for the difference between the two latter periods (EModE and PDE) with \( p = 0.49 \). The slight increase in the PDE period shown in Figure 1 cannot be, therefore, considered statistically significant.

Figure 1. Overall syntheticity by periods and word-classes.

<table>
<thead>
<tr>
<th>Word</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
<th>Word</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thing</td>
<td>0.47</td>
<td>0.41</td>
<td>0.30</td>
<td>0.30</td>
<td>Open</td>
<td>0.72</td>
<td>0.22</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Earth</td>
<td>0.43</td>
<td>0.16</td>
<td>0.02</td>
<td>0.01</td>
<td>Wise</td>
<td>0.89</td>
<td>0.35</td>
<td>0.22</td>
<td>0.34</td>
</tr>
<tr>
<td>Gold</td>
<td>0.51</td>
<td>0.23</td>
<td>0.04</td>
<td>0.02</td>
<td>Full</td>
<td>0.72</td>
<td>0.22</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Horse</td>
<td>0.62</td>
<td>0.41</td>
<td>0.28</td>
<td>0.17</td>
<td>Wide</td>
<td>0.63</td>
<td>0.30</td>
<td>0.17</td>
<td>0.28</td>
</tr>
<tr>
<td>Fox</td>
<td>0.38</td>
<td>0.44</td>
<td>0.29</td>
<td>0.19</td>
<td>Bright</td>
<td>0.89</td>
<td>0.58</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>Dust</td>
<td>0.40</td>
<td>0.14</td>
<td>0.00</td>
<td>0.01</td>
<td>Long</td>
<td>0.78</td>
<td>0.31</td>
<td>0.23</td>
<td>0.18</td>
</tr>
<tr>
<td>Oxa</td>
<td>0.44</td>
<td>0.47</td>
<td>0.31</td>
<td>0.31</td>
<td>High</td>
<td>0.84</td>
<td>0.50</td>
<td>0.28</td>
<td>0.37</td>
</tr>
<tr>
<td>Nama</td>
<td>0.27</td>
<td>0.18</td>
<td>0.20</td>
<td>0.24</td>
<td>Warm</td>
<td>0.67</td>
<td>0.35</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td>Assa</td>
<td>0.42</td>
<td>0.32</td>
<td>0.28</td>
<td>0.19</td>
<td>Hard</td>
<td>0.85</td>
<td>0.47</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>Lust</td>
<td>0.61</td>
<td>0.41</td>
<td>0.30</td>
<td>0.10</td>
<td>Warm</td>
<td>0.67</td>
<td>0.35</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td>Winter</td>
<td>0.45</td>
<td>0.37</td>
<td>0.09</td>
<td>0.08</td>
<td>Holy</td>
<td>0.69</td>
<td>0.39</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Hell</td>
<td>0.14</td>
<td>0.14</td>
<td>0.04</td>
<td>0.01</td>
<td>Grim</td>
<td>0.90</td>
<td>0.31</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>Ende</td>
<td>0.11</td>
<td>0.07</td>
<td>0.12</td>
<td>0.10</td>
<td>Great</td>
<td>0.77</td>
<td>0.39</td>
<td>0.24</td>
<td>0.34</td>
</tr>
<tr>
<td>Rest</td>
<td>0.15</td>
<td>0.27</td>
<td>0.00</td>
<td>0.02</td>
<td>Send</td>
<td>0.53</td>
<td>0.50</td>
<td>0.36</td>
<td>0.45</td>
</tr>
</tbody>
</table>
7. Conclusions and paths for further research

That the textbook story of a typological re-shaping in the history of English seems not to be in need of a thorough revision may be found reassuring, but this was neither a principal aim nor is it a primary conclusion of our probe. Much more importantly than that, we hope to have presented here another corpus-based methodology of measuring the analyticity-syntheticity dichotomy in a way that makes it possible to obtain precise, systematic, balanced and holistic measurements.

Our methodology also allows us to draw further conclusions besides the general trends of syntheticity. While calculating the standard deviation for the syntheticity of the individual words, we have noted different trends for different word-classes, namely, that the verbs and adjectives are becoming less homogeneous in their syntheticity levels, while nouns have become slightly more homogenous (Figure 2). What is especially noticeable is the rise of heterogeneity in the verbal subsystem after the OE period – some verbs lose most of their syntheticity, like let, while verbs like stand or help briefly increase their syntheticity levels (Table 6). The small increase in the heterogeneity of adjectives in the PDE sample is mostly caused by an increased frequency of comparative and superlative forms in some adjectives, while adjectives like dead or open exhibit very little or no synthesis at all (Table 7). In conjunction with the widening gap between the syntheticity levels of individual word-classes (Figure 1), we

<table>
<thead>
<tr>
<th>Word</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
<th>Word</th>
<th>OE</th>
<th>ME</th>
<th>EModE</th>
<th>PDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Womb</td>
<td>0.29</td>
<td>0.07</td>
<td>0.12</td>
<td>0.06</td>
<td>Sell</td>
<td>0.73</td>
<td>0.72</td>
<td>0.41</td>
<td>0.52</td>
</tr>
<tr>
<td>Hunger</td>
<td>0.40</td>
<td>0.11</td>
<td>0.00</td>
<td>0.02</td>
<td>Write</td>
<td>0.78</td>
<td>0.66</td>
<td>0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>King</td>
<td>0.40</td>
<td>0.39</td>
<td>0.22</td>
<td>0.14</td>
<td>Help</td>
<td>0.75</td>
<td>0.80</td>
<td>0.53</td>
<td>0.43</td>
</tr>
<tr>
<td>Dead</td>
<td>0.75</td>
<td>0.30</td>
<td>0.00</td>
<td>0.00</td>
<td>Stand</td>
<td>0.81</td>
<td>0.91</td>
<td>0.68</td>
<td>0.56</td>
</tr>
<tr>
<td>Dumb</td>
<td>0.77</td>
<td>0.28</td>
<td>0.00</td>
<td>0.01</td>
<td>Let</td>
<td>0.76</td>
<td>0.35</td>
<td>0.09</td>
<td>0.17</td>
</tr>
<tr>
<td>Warm</td>
<td>0.67</td>
<td>0.35</td>
<td>0.09</td>
<td>0.16</td>
<td>Bring</td>
<td>0.78</td>
<td>0.69</td>
<td>0.43</td>
<td>0.51</td>
</tr>
<tr>
<td>Open</td>
<td>0.72</td>
<td>0.22</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Overall syntheticity index per word and period.
may conclude that although the inflectional system of the language as a whole is becoming less heterogeneous, some of its subsystems, like verbs or adjectives, have become – mainly due to the nature of the items in our samples – more heterogeneous. We believe this to be a signal of a deteriorating inflectional system, a system which is in general less synthetic but keeps certain isolated vestiges of the older system intact (irregular verbs, plural nouns, adjectives in comparative and superlative), or even reinforces their position (verbal -ing forms) (Tables 2-4).

Figure 2. Standard deviation (SD) of individual words’ syntheticity levels calculated for each word-class and in total.

<table>
<thead>
<tr>
<th></th>
<th>STAND</th>
<th>LET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OE</td>
<td>ME</td>
</tr>
<tr>
<td>-0</td>
<td>2.75%</td>
<td>6.69%</td>
</tr>
<tr>
<td>-te</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-e</td>
<td>16.25%</td>
<td>17.39%</td>
</tr>
<tr>
<td>-(e)th/s</td>
<td>5.95%</td>
<td>13.04%</td>
</tr>
<tr>
<td>-ed</td>
<td>0.00%</td>
<td>0.33%</td>
</tr>
<tr>
<td>-en</td>
<td>0.69%</td>
<td>13.21%</td>
</tr>
<tr>
<td>-ing</td>
<td>0.00%</td>
<td>5.85%</td>
</tr>
<tr>
<td>-V3-0</td>
<td>0.00%</td>
<td>20.74%</td>
</tr>
<tr>
<td>-V4-e</td>
<td>1.60%</td>
<td>15.38%</td>
</tr>
<tr>
<td>-an</td>
<td>35.24%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ath</td>
<td>13.96%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-don</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-V2-en</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ton</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Measuring typological syntheticity of English diachronically with corpora

<table>
<thead>
<tr>
<th>STAND</th>
<th>ME</th>
<th>LET</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>-V4-on</td>
<td>7.78%</td>
<td>0.17%</td>
<td>4.71%</td>
</tr>
<tr>
<td>-ande</td>
<td>14.42%</td>
<td>3.18%</td>
<td>0.23%</td>
</tr>
<tr>
<td>-st</td>
<td>0.46%</td>
<td>2.34%</td>
<td>1.03%</td>
</tr>
<tr>
<td>-edest</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-anne</td>
<td>0.23%</td>
<td>0.00%</td>
<td>0.57%</td>
</tr>
<tr>
<td>-V4-en</td>
<td>0.69%</td>
<td>1.67%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-den</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>-test</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SYNTHETICITY</td>
<td>0.81</td>
<td>0.91</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 6. Increasing heterogeneity of syntheticity levels exemplified by the verbs *stand* and *let* in OE and ME.

<table>
<thead>
<tr>
<th>HIGH</th>
<th>OPEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EModE</td>
<td>PDE</td>
</tr>
<tr>
<td>-0</td>
<td>79.40%</td>
</tr>
<tr>
<td>-e</td>
<td>9.55%</td>
</tr>
<tr>
<td>-ra</td>
<td>11.05%</td>
</tr>
<tr>
<td>-an</td>
<td>0.00%</td>
</tr>
<tr>
<td>-st-an</td>
<td>0.00%</td>
</tr>
<tr>
<td>-es</td>
<td>0.00%</td>
</tr>
<tr>
<td>-r-an</td>
<td>0.00%</td>
</tr>
<tr>
<td>-ena</td>
<td>0.00%</td>
</tr>
<tr>
<td>-u</td>
<td>0.00%</td>
</tr>
<tr>
<td>-st-um</td>
<td>0.00%</td>
</tr>
<tr>
<td>-r-um</td>
<td>0.00%</td>
</tr>
<tr>
<td>SYNTHETICITY</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 7. Increasing heterogeneity of syntheticity levels exemplified by the adjectives *high* and *open*.

Of course, the validity of our conclusions must be further tested against a wider empirical base, namely by complementing syntheticity measurements by those of analyticity. Moreover, and just as importantly from a methodological viewpoint, interpreting the slight increase in syntheticity between EModE and PDE (as cumulatively
shown in Figure 1), however statistically insignificant in itself, reaffirms the importance of typological theory for analyses of this kind. This slight increase may well be due to a large presence in the corpus data of such markers as PDE -ing, naturally construed, in terms of quantitative typology, as synthetic. The role of this marker may, however, be typologically much more complex than meets the eye. On the one hand, it is customarily associated with condensation structures of an analytical (or, in terms of Prague School Typology, *isolating*) syntax. On the other hand, its systemic equivalents, i.e., functionally comparable vehicles of sentence condensation, can be also shown to operate in syntactic systems that are predominantly agglutinating in nature. In other words, not only must syntheticity measurements be balanced by those of analyticity, but also quantitative typology should be considered along with a ‘qualitative’, holistic typological assessment of language systems, i.e., such as would, in a detailed and balanced way, take account of ‘inflectional’ and ‘agglutinating’ patterns of a broadly conceived typological ‘syntheticity’.

References


*PPCMBE* = Kroch et al. (2010)


YCOE = Taylor et al. (2003).
Part 2

Corpora and Descriptive Linguistics
Abstract

Dictionaries are sometimes used as sources of data for similar purposes as corpora. The value of some of the evidence retrievable from dictionaries is questionable compared with the value given to similar evidence retrieved from corpora. This question can be raised comparing the contents of this part of the book and also of some of the contents of the first part. In both cases, the data obtained from corpora and dictionaries advise to revise statements that are according to logical expectations or according to what is in the literature.

1. Introduction

In a forthcoming paper, Laurie Bauer and myself point out the existence of a semantic pattern in English noun/verb conversion that diverges interestingly from the ones described in the literature (e.g., Marchand 1969: 365, Aronoff 1984: 45-46, Plag 1999: 219-220, among others). The paper is based on the analysis of every instance of noun/verb conversion (as described in synchronic terms) contained in the *Oxford English Dictionary (OED)* (Simpson/Weiner 1992). Although the evidence found is relevant, it is little, especially compared with the number of examples that follow the usual patterns. The

---

1 Author’s email address: svalera@ugr.es; affiliation: U. of Granada. I would like to thank Ana Díaz-Negrillo, Alejandro Alcaraz-Sintes and Juan Santana-Lario for helpful comments.
question arises whether this semantic pattern is relevant enough as to be added to the others in the literature and which account for the majority of instances of noun/verb conversion. A similar study on noun/verb conversion in English based on evidence obtained from the *British National Corpus (BNC)* supports this as regards semantic patterns in noun/verb conversion and the pattern mentioned above.

The fact that the dictionary and the corpus confirm the same results raises the following question on the use of dictionaries as bodies of data: What are the differences between dictionaries and corpora, when both are used as sources of data for the same research question? In other words, what are the differences between the conclusions drawn from dictionaries and the conclusions drawn from corpora, if both are used as sources of data for the same question?

Regarding the value of the findings of dictionary- vs. corpus-based research, whereas dictionaries contain interpretations of data, computerised corpora contain raw data, or data supported by annotation. Lexicographical practice relies on lexicological principles that are not always easy to apply, and encodes information according to a system selected by dictionary builders to meet general and specific uses. Differences may be noticed between dictionaries as regards, e.g., sense separation, because the interpretation of data differs, or because the system used for the presentation of information differs. This is probably enough to discard the former in order to avoid potential bias. Less commonly, but not infrequently, dictionaries are discarded on the grounds that they are more about lexical structure than about other levels of description. In the case of noun/verb conversion, there is also the difference that dictionaries do not offer the quantitative data that corpora contain as regards frequency, genre distribution and other issues, for example word combinations. Frequency is essential because it allows to draw conclusions when it is high and low, especially when the frequency in a large corpus is 1. Qualitatively, and except for large dictionaries like the *OED*, dictionaries rarely provide such comparable contexts or allow overviews of the range of use, and this applies also to the case of noun/verb conversion.

These are not the only differences between dictionaries and computerised corpora, but they help us focus on an interesting point (for a better review of dictionaries as sources of data compared with
computerised corpora, cf. Mair 2004). In the paper on conversion that we are considering here, a comprehensive dictionary was preferred over a corpus because the dictionary contains, by definition, a qualified semantic interpretation of the two terms under research, noun and verb. The use of a corpus instead of a dictionary as a source of data refers back to a dictionary anyway for the analysis of each word, if the aim is to identify the semantic contrast between the two categories. Use of a dictionary from the start disposed of a stage in data processing, namely the interpretation of the semantic contrast between the noun and the verb, at the cost of variables like frequency or distribution.

Another difference is that, at least in the case that we are considering here, that is, the existence or not of a specific formation, it can be assumed that the dictionary contains evidence of sanctioned use (regardless of the authority or of the value granted to the sanction implicit in the incorporation of an item to a specific dictionary or to dictionaries in general). By contrast, the corpus contains evidence of use. This means that attestation of only one instance of a particular formation or a particular case in a dictionary has at least the value of having been assessed as wider in use than a nonce-formation or, if it is recorded in a dictionary as a nonce-formation, it has the value that it has been judged to be relevant enough as to be included in the dictionary (again, regardless of the measurement instruments used for the assessment of the frequency or of the relevance of the item in question, and of their validity or of their legitimacy over time). Admittedly, and unless based on corpora, the dictionaries’ assessments of the usage profile of items (for example, as rare, obsolete or, more often, neither of these when in fact they have fallen into disuse) have been called into question, whereas occurrences in computerised corpora are evidence of use, with the reservation of hapax legomena.

Hapax legomena are attestations of only one case in a corpus. They have become a standard measurement index for morphological productivity, even if single occurrences may signal differently than new formations (Baayen & Lieber 1991, Baayen 1992a, 1992b, Plag 1999, Bauer 2001; cf. also Pustylnikov & Schneider-Wiejowski 2010 for a discussion of what hapaxes may represent besides neologisms).
The possibility of a single occurrence in a large dictionary, for example of an affix, a syntactic or a semantic pattern, addresses in part the question of what are the differences between the conclusions drawn from dictionary evidence and those drawn from corpus evidence. In this particular case, and in the context of the study of noun/verb conversion, the possibility of a single occurrence of one pattern or of one case in a dictionary suggests that the value or the interpretation of its singularity is a major difference (considering that these single occurrences are also instances of one-case records and that, unlike hapax legomena, dictionary entries have been assessed as in use by lexicographical practice, which disposes of some of the interpretations that hapax legomena lend themselves to; again, cf. Pustylnikov & Schneider-Wiejowski 2010). The following section reviews the papers of Part 2 of this volume in case any of them may present a similar case to the one posited in the introduction and how it has been dealt with.

2. Dictionary- and corpus-based research in applied and descriptive linguistics

The use of dictionaries as sources of data in the way of corpora is frequent, even if it is largely constrained by the research object, and even if it is not so frequent in applied linguistics. This is exemplified in this volume by two papers in the field of applied linguistics and five in descriptive linguistics.

Starting with applied linguistics, the papers contained in the second part of this volume illustrate the above. The two papers that deal with research learner language (Meurers/Krivanek/Bykh’s and Thompson’s) are examples of corpus use for the development of corpus tools or as descriptions of corpus evidence. In this volume Meurers/Krivanek/Bykh use two experiments in native language identification. The experiments explore automatic linguistic annotation of non-native language and linguistic modelling in the interaction of specific surface
forms and broader, abstract categories. One of the experiments by Meurers/Krivanek/Bykh tests the application of part-of-speech categories based on the *International Corpus of Learner English (ICLE v. 2)*, and the other tests syntactic verb alternation.

The implications of these results gear further understanding both of linguistic theories in general and of language acquisition. The conclusions prove the need for multi-level annotation in learner corpora if theoretically consistent research is aimed at annotated learner language. Difficulties lie in striking the right balance between attention to detail in the surface structure and the abstract categories to which the elements in the surface respond.

The second paper of this type, the one by Thompson, uses Hoey’s (2005) theory of lexical priming as a framework for his study. Specifically, the author focuses on a less frequent research issue, textual colligation, in non-native language as evidenced in essays contained in the *British Academic Written English* corpus (*BAWE*). Despite the reservations of the author and also despite the inherent diversity across different types of essays in the corpus, the method stands out for the selection of the lemmas used in the study according to their occurrence in paragraph-initial and paragraph-final sentences of texts.

The paper researches the distribution of such dissimilar language items as the noun *essay*, the adjective *important*, the linking adverbs *thus* and *therefore*, phrases of the type *it is * to, *be able to, this essay, one of the *, and modal verbs. The analysis of the occurrence of specific words according to discipline and genre, and then of any differences between various levels of language proficiency reveals preferences that support the notion of textual colligation. Importantly, the paper contributes two ideas that are more telling than the findings about textual colligations: primings are tied to *rhetorical* contexts, and priming is refined and becomes established from experience.

The papers of linguistic description are interestingly different from the former two papers as regards their use of dictionaries or of computerised corpora in Part 1 and in Part 2 of this volume. In the diachronic part, Part 1, Calvo-Cortés and then Fanego use dictionaries as sources of data. Their research illustrates some of the advantages and of the disadvantages mentioned in the *Introduction*. In Calvo-Cortés,
the data from historical corpora partly disprove the evidence contained in the *OED*. In Fanego, historical dictionaries and thesauruses are used for attestation of existing forms, which are then quantified and related to their ultimate language sources and to their formation mechanisms (semantic extension vs. derivation), and whose historical evolution, lexical meaning and syntax is then analyzed. Fanego’s second type of conclusions, the methodological ones, are particularly relevant in that at least certain aspects of dictionaries as sources of data are considered at the same level of relevance as historical corpora.

By contrast, none of the chapters of Part 2 use dictionaries as primary sources of data. This is partly because their research topics rely on evidence that is usually not contained in dictionaries. This is the case of four out of the five remaining papers in this Part: they research structures and their syntactic, semantic and textual features are not recorded lexicographically.

Benítez-Castro researches the concept ‘shell noun’ based on the evidence provided by the *BNC Sampler* according to nine variables that follow Hoey’s (2005) theory of lexical priming and bring into the analysis all levels of linguistic description and therefore fall beyond the evidence that dictionaries usually supply. The aim is a comprehensive description of what makes a noun qualify as a shell noun, as described in the literature.

Unlike previous approaches to shell nouns, Benítez-Castro uses a manual and context-sensitive approach to the corpus. This is important for two reasons: first, because his choice of a general or genre-balanced corpus sets his paper apart from other corpus studies of this issue, which are mainly genre-specific; second, because replacement of the automatic retrieval of examples of shell nouns based on structural patterns with manual retrieval from a lexical list based on the literature ensures representation of a wider range of instances than those contained in the structural patterns used for automatic retrieval.

Regardless of the conclusions drawn about shell-nounhood, the paper underlines the relevance of manual corpus analysis for comprehensive results, even if this means the use of fewer examples (because this approach means more time and effort). By way of example, in the specific case of shell nouns, this approach contributes, among others, results that underline the importance of prepositional phrases to the
detriment of noun complement clauses, as well as the importance of cataphor ic reference compared with that of anaphoric reference.

The chapter by Coto-Villalibre is a study of the passive voice with focus on get-passives. The paper is an analysis of their syntactic, semantic and pragmatic features, a comparison with be-passives, and then a review of the classification of get-passives as central, semi-, pseudo-, adjectival, idiomatic and reflexive constructions. The source of data is a relatively small computerised corpus, the spoken part of the British module of the International Corpus of English.

The results confirm differences between the types of get-passive structures established and, more interestingly, they deviate from the expectations collected in the literature as regards the occurrence of the feature [+ANIMATE] in the by-agents of get-passives (lower than reported), and also as regards the occurrence of responsible subjects in get-passives and semi-get-constructions (lower than expected). These findings prompt the question of whether the discrepancies found are specific to the corpus used (spoken British English), whether they are a general tendency independent of register or language variety, and of what other unexpected results may appear for related structures (or not) across registers, genres and language varieties.

Santana-Lario researches adjective complementation by whether/if-clauses based on the British National Corpus. The paper is of the same type as Benítez-Castro’s, Coto-Villalibre’s and Mato-Míguez’s in that the research question is about a structure and does not lend itself easily to analysis based on dictionary evidence. After automatic retrieval of a sample of adjectives complemented by this type of clauses, Santana-Lario annotates the examples obtained manually for three main variables: syntactic construction, form and meaning of the adjectival head, and contextual features such as the co-occurrence of negative or non-assertive elements and alternative coordination with or.

The results show relevant differences in the selection of whether or if according to the position of the adjective, such that whether is preferred with predicative adjectives. By contrast, the results show similarities between whether and if clauses in the occurrence of a semantic feature of negativeness or non-assertiveness throughout the adjectives that take complementation by these two clauses, and in the
expression of alternates. These results are important for their divergence from the descriptions available in the literature where, for example, these complement clauses are viewed as interrogative in nature. More important, the paper lays new emphasis on the complexity of the relation between lexis and its grammatical complements and the influence of general semantic features on specific realizations. The results presented in this paper sound out the intricacy of these relations and, by doing so, they reinforce the need for further corpus research on them.

Thus, only one of the papers of Part 2 could have used dictionaries as sources of data, and even this in addition to (not instead of) computerised corpora, because replacement of the quantitative data for this paper via a dictionary would have required a different and probably costlier strategy than via corpora. The chapter by Egan contrasts how the concept ‘throughness’, whether literal or figurative, is encoded in English and French as compared with their Norwegian correlate preposition/particle gjennom, as recorded in a parallel corpus, the Oslo Multilingual Corpus. The occurrences of the original Norwegian term are classified semantically as eight semantic categories and their counterparts are then expressed either as prepositions or not recorded in English and French in the same study corpus.

The results reveal statistically significant differences in the way English and French encode the notion of ‘transfer’ or ‘going through’, with a French tendency towards use of prepositions that is considerably higher than the English one and that is in line with the results obtained for other concepts elsewhere. The use of what may be assumed to be the default English and French prepositions for Norwegian gjennom (through and à travers, respectively) reveal a small overlap between the two languages. By semantic classes, English and French tend to coincide more in the sense of perception ‘throughness’ than in the sense of motion ‘throughness’, in the latter case for typological reasons. The paper is particularly interesting in that it shows that the potential for further research in this field is enormous.
3. Conclusions

Language research avails itself of a variety of resources. Computerised corpora stand as the most reliable and widespread empirical source of data. In this volume use of dictionaries as sources of data is also exemplified in papers where the research topic lends itself to it, that is, research in applied and descriptive linguistics uses dictionaries and computerised corpora according to their objects of study. In the papers contained in this volume dictionaries are used less frequently as sources of data, largely because the research topics require access to information that is often not contained in dictionaries.

This does not help answer the question asked at the beginning of this chapter, specifically about the value of limited evidence in a dictionary and in a computerised corpus, but in any case this is a research topic in itself and largely an open question for the research community. It would have been interesting to see how the papers which use dictionaries would react to certain types of evidence, that is, whether the same value is granted to certain types of apparently similar quantitative evidence contained in dictionaries as when it is contained in corpora.

By contrast, what can be clearly seen in these papers is that both dictionaries and computerised corpora keep bringing to light data that are not in accordance with, or which simply are against logical expectations or against what the specialized literature has assumed to be the case in specific areas. In some cases this needs confirmation through the use of more comprehensive corpora, but in most others it is further proof of how enlightening dictionary- and corpus-based research is.
References


Abstract

This paper analyses a group of shell nouns from a small but representative sample of the English language, i.e., the BNC Sampler Corpus. This class of nouns comprises abstract units (e.g., lie, idea, issue) that help to condense long stretches of discourse into smaller discourse entities. The motivation for their use lies in their text-organising and evaluative functions. The automation and genre-specific nature of most research on these units provides only a partial account of their use in natural discourse. This article therefore offers a tentative profile of shell-noun behaviour based on a manual and context-sensitive analysis of nine variables covering all levels of linguistic analysis (i.e., formal, syntactic, semantico-pragmatic and textual). The evidence (42 shell nouns: 1,110 concordance lines) reveals certain similarities with the findings in (fully and semi-) automated analyses of the literature (e.g., genre: academic and journalistic; deictic: specific), but also various differences (e.g., syntactic pattern: noun complement clauses vs. prepositional phrases; rhetorical function: more anaphoric vs. more cataphoric).

Author’s email address: mabenitez@ugr.es; affiliation: U. of Granada. This paper has been funded by the Spanish Ministry of Education (research grant FPU AP2008-03204). Special thanks are owed to the two anonymous referees and (in alphabetical order) to Leanne Victoria Bartley, Encarnación Hidalgo Tenorio and Paul Thompson.
1. Introduction

Numerous studies have cast light on the role of abstract nouns like \textit{objective} or \textit{problem} in the rhetorical organisation of certain broad written genres, namely academic prose and newspaper language. This raises the question of whether the encapsulating and reifying discourse function of these nouns, labelled here as \textit{shell nouns} (Hunston/Francis 2000; Schmid 2000), is genre-specific or a wide-ranging phenomenon of language. Closely linked to the corpus employed is the type of analysis adopted, ranging from the fully automated analysis of a large corpus such as the \textit{Bank of English (BoE)}, e.g., Schmid (2000) to the manual analysis of a small corpus, e.g., Flowerdew (2003). The former relies on a set of pre-defined queries modelled on the syntactic patterns assumed to prevail in the contextual use of these nouns, i.e., ‘N-Cl’ and ‘N-be-Cl’. Manual analyses are often centred on genre-specific corpora (e.g., Flowerdew 2003: biology lectures vs. textbooks) and appear to be mostly concerned with the identification of the rhetorical features of these nouns. Thus, there seems to be a need for a study that employs a small though more balanced corpus in order to provide an all-encompassing linguistic perspective on the use of these units.

This paper aims at the manual identification of formal, syntactic, textual and semantico-pragmatic features of shell-noun phrases. The goal is to provide an overall account and tentative profile of shell-noun behaviour based on a manual and contextualized analysis of a representative sample of the English language at large, i.e., the \textit{British National Corpus Sampler (BNC Sampler)}. The analysis draws on a list of 950 lemmas extracted from the literature. The evidence studied for this paper consists of 42 shell nouns from three frequency ranges of the aforementioned list (i.e., top, middle and bottom). A random set of 40 concordances per lemma is drawn from the corpus. In view of the potential priming differences between word-forms (Sinclair 1991), 20 concordances are analysed for the singular and 20 for the plural.
The article first offers an overview of the key areas involved in the description of these units and the corpus methods and procedures employed in related research (§2). The manual corpus-driven methodology of this study is then presented (§3). Finally, a set of generalisations about shell noun behaviour in the literature are compared with the findings from the present study (§4).

2. Theoretical background

2.1. Definitions

The class of nouns which, in line with Hunston/Francis (2000) and Schmid (2000), shall be referred to here as *shell nouns* has been the subject of considerable academic debate and scholarly research over the past four decades. As a result, numerous terms and definitions have been proposed to account for the differing views on the scope of “[t]he property of shell-nounhood” (Schmid 2000: 13). In broad terms, such a property may be assigned to nouns which “[...] *talk about* the ongoing discourse” (Francis 1986: 3), which “[...] encapsulate or package a stretch of discourse” (Francis 1994: 85) and whose meaning is twofold, i.e., specific and dictionary (Ivanič 1991: 95). Such discourse features tend to manifest themselves in head nouns “accompanied by the reference item *the*” (Halliday/Hasan 1976: 275) and “[...] followed by a *that*-clause [...]” (Hunston/Francis 2000: 186). Example (1) below shows a prototypical example of these nouns:

(1) “Mr Bush said Iraq’s leaders had to face the **fact** that the rest of the world was **against them**” (Schmid 2000: 22)

---

2 For the sake of clarity, the typographical conventions found in the presentation of examples in the literature will not be followed here. In all the examples cited in this paper (those from the literature and those from the present study), boldface highlights the shell-noun head and underlining identifies the antecedent in each case.
In the literature to date, the description of these units rests on nine different terms, all of them bringing to the forefront their discursive functions, i.e., *general nouns* (Halliday/Hasan 1976: 274-282; Mahlberg 2005), *Vocabulary 3 items* (Winter 1977), *lexical signalling* (Hoey 1979), *enumerables* and *advance labelling* (Tadros 1985, 1994), *anaphoric nouns* (Francis 1986), *carrier nouns* (Ivanič 1991), *advance and retrospective labels* (Francis 1994), *shell nouns* (Hunston/Francis 2000; Schmid 2000) and *signalling nouns* (Flowerdew 2003). Certain similarities and differences are evidenced in such categorisations, those of which may be summarised in the following.

A key factor frequently adduced in the description of shell nouns is that of abstraction. This notion relies on formal and semantic criteria. Formal criteria lie at the core of Vendler’s (1968) approach to the ontological classification of nominals (Events or Facts), based on their appearance in fixed structural patterns known as “containers” (Vendler 1968: 33). One such container is the “N is N” pattern, where the shell noun occurs as complement of copular be (see (2) below). Winter’s (1977: 9, 85) discussion of ‘Vocabulary 3’ items is also dependent on form, as illustrated by his claim that only prototypical concrete nouns (e.g., *animal*) would prove felicitous in an example like (3) below. Lyons’ (1977: 442-445) description of abstraction is more semantic in nature, insofar as no connection is established between the ontological status of nominal units and typical sentential patterns. His threefold semantic classification of nouns comprises “first-order entities”, “second-order entities” and “third-order entities”. The former would correspond to prototypical concrete nouns (e.g., *table, spoon*, etc.), whilst the latter would comprise nominalised processes and events (e.g., *activity, destruction*), and facts and propositions (e.g., *issue, problem, argument*), respectively. Schmid (2000: 68) argues that in a scale of prototypicality, third order entities constitute the “core of the class of shell nouns”.

(2) “[That he died]/[His death] is a fact” (Vendler 1968: 73)

(3) “It is like an effect […]” (Winter 1977: 85)
Another distinguishing feature of these units concerns the often observed indeterminacy between open-class and closed-class items (e.g., Halliday/Hasan 1976: 275; Winter 1977: 2; Francis 1986: 3; Ivanič 1991: 103). Their open-class status stems from their “dictionary” meaning, while their closed-class status relates to their “specific” or context-dependent meaning (Ivanič 1991: 95). Shell nouns are thus argued to carry an inherent “dictionary” meaning, remaining constant, and one which, like pronouns, varies depending on the context where the noun appears. Such a variable meaning may be explained by reference to Lyons’ (1977: 668) concept of “impure textual deixis” and Fraurud’s (1992) notion of “situation reference”. Both terms capture the idea that, in some cases of anaphoric reference, the pronoun is not co-referential with a first-order entity, but with a longer stretch of discourse. Example (4) offers a typical instance of the more concrete or “object” (Faururd 1992: 3) type of anaphora, whilst (5a) and (5b) are instances of the more impure or situation-dependent anaphora. Examples (5a) and (5b) contain the same information, the only difference lying in the use of a referential demonstrative pronoun in (5b) and a referential shell-noun phrase in (5a). The use of a shell noun in (5a) endows the writer with a means of evaluating the underlined stretch of discourse. It is this “characterisation” potential of shell nouns (Schmid 2000: 13) that brings them closer to the open-class (lexical) end of the continuum.

(4) “The boys went home. They were tired” (Faururd 1992: 3)
(5a) “[…] the supply of children far exceeds the pool of people prepared to provide them with a home and a family life. The problem was highlighted this week when a newspaper in Oxford published […]” (Schmid 2000: 350)
(5b) “[…] the supply of children far exceeds the pool of people prepared to provide them with a home and a family life. This was highlighted this week when a newspaper in Oxford published […]”

From the examples presented so far, the tendency seems to be for shell nouns to refer to or “encapsulate” (Sinclair 1993: 7) stretches of discourse. This implies that, in principle, noun-to-noun or “object” (Faururd 1992: 3) reference should be no guarantee of “shell-nounhood”. The literature consulted appears to support this argument. Francis (1994: 85), for example, categorically states that her “major
criterion for identifying an anaphorically cohesive nominal group as a retrospective label is that there is no single nominal group to which it refers”. Some references (Ivanič 1991: 109; Flowerdew 2003: 336; Gray 2010: 179), however, contemplate the possibility of noun phrase antecedents for shell-noun phrases, as illustrated in (6):

(6) “It is interesting to read about the items electors mentioned as having, in their view, specially affected the election. […] rash Labour promises – cost of new pension scheme – bribery of electorate […] strikes” (Ivanič 1991: 109)

When identifying antecedents or “shell-content” (Schmid 2000: 9), attention is also paid to the referential status of noun phrases in discourse. The treatment of reference in the study of shell nouns is inspired by the extensive literature on (mostly) pronominal anaphora resolution (e.g., Chomsky 1981; Fraurud 1988; Asher 1993). As such, it is unsurprising that most research focuses only on anaphoric shell-noun use (e.g., Halliday/Hasan 1976; Francis 1986; Conte 1996; Charles 2003; Moreno 2004; Gray 2010). The literature reviewed reveals comparatively few studies on other types of reference. Some references focus only on the cataphoric function of these items, as shown for example in Winter (1977), Tadros (1985), Hunston/Francis (2000) and Charles (2007). The interplay between anaphora and cataphora is explored in Francis (1994), Mahlberg (2005) and Schmid (2000). Finally, both endophoric (anaphoric and cataphoric) and exophoric uses are mentioned in Ivanič (1991), Partington (1998) and Flowerdew (2003). One caveat here is that even when an all-encompassing approach to reference is adopted, no quantitative data are offered about the frequency of different types of reference.

In relation to the link between antecedent and noun, the literature consulted appears to be primarily concerned with the intersentential functions of shell nouns. Comparatively few studies mention reference either within the boundaries of the sentence (e.g., Hunston/Francis 2000; Biber 2006; Charles 2007) or both within and outside these boundaries (e.g., Ivanič 1991; Winter 1992; Schmid 2000; Flowerdew 2003; Caldwell 2009). The foregoing discussion has brought to light the lack of consensus apparent in much related research in respect of the treatment of reference. Drawing on Winter
(1982: 32), therefore, a need arises for a comprehensive approach to the study of signalling in discourse, to the extent that emphasis should be laid on “the signalling role of any word in the clause, whether it signals backwards in its sentence or beyond its sentence to a preceding sentence, or whether it signals forward within its sentence or beyond its sentence to a sentence which follows it”. The three examples below illustrate intersentential anaphora (7), intrasentential cataphora (8) and exophora (9):

(7) “[…] the Soviet Union has ‘shot its bolt’, and that only the unreconstructed Cold Warriors are losing any sleep about the Russian menace. // James Reston has readily and complacently echoed this assessment in his criticism of the Reagan équipe” (Francis 1986: 27)

(8) “It is my contention that pressure from parts of the Labour Party and Movement was not central to the decision taken to withdraw” (Charles 2007: 208)

(9) “[…] the associations can call on the resources of the Commonwealth Mycological Institute which maintains a collection of fungi many of which are of interest in research into certain food problems” (Ivanič 1991: 106)

Finally, as regards the formal structure of the noun phrase, shell nouns are often claimed to occur only with so-called “specific deictics” (Halliday/Matthiessen 2004: 312-317), i.e., the definite article, demonstrative determiners and possessive determiners. Only few references allow for the presence of the indefinite article in shell-noun use (Quirk et al. 1985: 1261; Ivanič 1991: 111; Partington 1998: 92-93; Aktas/Cortés 2008: 10) (see (10)).

(10) “In a move to tighten control of a far-reaching empire and to improve the group’s own image, Maurice and Charles Saatchi […] have stepped down from the day-to-day running of the group” (Partington 1998: 94)

The literature also places considerable emphasis on two post-head structures, i.e., that- and to-infinitive “noun complement clauses” (Biber et al. 1999: 645), as shown in NOUN + CLAUSE (see (11)) and NOUN + BE + CLAUSE patterns (see (12)). Prepositional phrases, paramount among the types of noun phrase post-modification according to Biber et al. (1999: 634), tend to be disregarded in most shell-noun descriptions available. Some exceptions (see (13)) include

(11) “The Association will give a warning that poll tax bills in some Conservative districts will exceed government guidelines […]” (Schmid 2000: 135)

(12) “The first action was to place the vessel under cover and remove the deck-house” (Schmid 2000: 263)

(13) “its function of providing mechanical strength” (Flowerdew 2003: 337)

2.2. Corpus methodology and analytical procedures

This section addresses the corpus methodology followed in the literature on shell nouns. Two issues will be dealt with here, one being the corpora employed and, the other, the kind of procedure adopted for the analysis of such corpora. Most of the generalisations made thus far about the use of shell nouns (see §2.1) stem from the analysis of corpora which are both small and genre-specific. One genre, academic discourse (both written and spoken), occupies a prominent role in most related research. Four subtypes of academic corpora may be distinguished, i.e., discipline-specific professional writing (as found in textbooks, research articles and theses: e.g., Tadros 1985; Moreno 2004; Charles 2007; Gray 2010), undergraduate L2 learner writing (e.g., Francis 1988; Flowerdew 2006, 2010; Hasselgård 2012), L1 (usually professional) vs. L2 (undergraduate or graduate) writing (e.g., Aktas/Cortés 2008; Caldwell 2009) and lectures (e.g., Flowerdew 2003; Lorés 2006). Another genre featuring in the literature is journalistic prose, epitomised by Francis’ (1986, 1994) use of articles from The Times and the monthly journal Encounter. It should also be noted that some genre-specific research has relied on diachronic corpora, as in Kanté (2010), which compares a synchronic corpus of research papers on linguistics with a diachronic corpus of judicial proceedings.

Compared to the dominance of genre-specific references, very few studies to date have made use of general English corpora. The three corpora employed comprise The Lancaster-Oslo/Bergen Corpus (LOB) (Leech/Johansson 1976; see Ivanič 1991), The Bank of English
(BoE) (through its different compilation stages: Hoey 1993; Hunston/ Francis 2000; Schmid 2000; Mahlberg 2005) and The British National Corpus (BNC) (Aijmer 2007; Yamasaki 2008). LOB, containing about 1 million words of only written English, is by far the smallest of the three corpora. Ivanič (1991) utilizes this corpus as a source of examples, failing to provide any frequency data about genre-related distribution of units and patterns. The BoE, amounting to 450 million words by the time Mahlberg (2005) went into print, is the largest of the general English corpora available. However, as Schmid (2000: 43) recognizes, this corpus provides size, but not balance, inasmuch as “more than two thirds of the whole material” correspond to newspaper language. Lastly, the BNC, with about 100 million words, is not as large as the BoE, but in terms of genre distribution, it is much more balanced than the BoE (notwithstanding the 90% vs. 10% proportion of written and spoken modes). The two studies where the BNC is used, nevertheless, do not report any genre-specific findings, insofar as the focus is placed solely on mode (i.e., spoken language: Aijmer 2007, and written vs. spoken language: Yamasaki 2008).

From the above it follows that the choice of a more or less genre-balanced corpus may play a significant role in the results obtained. Using a small genre-specific corpus will offer revealing insights into shell-noun behaviour in a specific type of written or spoken text, but, at the same time, one should be aware that, unless findings are compared with those from a more balanced corpus, no generalizations ought to be made about such behaviour in general English. Just as relevant to a description of shell-noun use is also the procedure applied to the analysis of the corpus. The literature consulted reveals three types of analysis, i.e., fully automated, semi-automated and manual.

Fully automated analyses (e.g., Francis 1993; Biber et al. 1999; Hunston/Francis 2000; Schmid 2000, 2007) are those conducted on large corpora (most commonly, the BoE). Automation here relates to the retrieval of examples from the corpus based on pre-defined queries, those of which typically represent the NOUN + CLAUSE and NOUN + BE + CLAUSE patterns (see §2.1). Schmid (2000), for example, compiled a list of 670 shell nouns after searching the BoE for these two patterns. The appearance of concrete nouns like book or
blackboard in examples such as this book and this is a blackboard led to the initial exclusion of the DEMONSTRATIVE/DETERMINER + NOUN and DEMONSTRATIVE + BE + NOUN patterns from the analysis. Shell-noun examples for these two patterns were subsequently retrieved on the basis of the 670 units previously identified after the initial clausal structure queries. The automation of this method thus reveals an inherent lack of flexibility resulting in “systematic misses” (Schmid 2000: 51-53). Such “misses” apply not only to exclusions of potential shell-noun instances occurring only in the DEMONSTRATIVE/DETERMINER + (BE) + NOUN pattern, but more importantly, to cases where the head noun and its lexical realisation (i.e., the complement clause) are interrupted by linguistic material, most frequently prepositional phrases. Example (14) exemplifies the exclusion of a prototypical shell-noun realisation based on the presence of an intervening prepositional phrase (i.e., of the project).

(14) “The next part of the project is to go back and to identify where these products come from” (Schmid 2000: 52).

A moderate degree of automation appears in those studies where predefined querying is followed by manual analysis. Two references illustrate this approach, namely, Yamasaki (2008) and Caldwell (2009), the former using the BNC and the latter using three small corpora of undergraduate L1, L2 and professional writing on cognitive psychology. Yamasaki (2008) looks at the two demonstrative patterns above (with and without be) in terms of their evaluative potential. The decision to focus on both follows the retrieval and subsequent manual analysis of a set of random concordances for 73 nouns. Queries for DEMONSTRATIVE + NOUN and DEMONSTRATIVE + BE + NOUN are then run on the corpus based on five nouns, i.e., change, shift, failure, mistake and problem. A subsequent manual analysis identifies the type of premodification and generic distribution (written or spoken English) of the nouns selected. Caldwell (2009: 78) conducts what she calls “Syntactically-motivated Searches” of both typical shell-noun patterns and patterns excluded from fully automated analyses (e.g., NOUN + PREPOSITIONAL PHRASE), with a view to comparing their
occurrence across non-professional native and non-native writing, and professional writing. A manual analysis then establishes the kind of reference that definite shell-noun phrases show in the first 200 words of every text that makes up the three corpora. From the methodology followed in Yamasaki (2008) and Caldwell (2009), their semi-automaticity is now evident, to the extent that predefined queries of patterns are used as a starting point for some sort of experimental manual analysis of contextualised data.

Of the three procedures, manual analysis constitutes the most corpus/data-driven approach to the study of shell nouns. In general terms, no predefined patterns are used and small genre-specific samples of language are preferred over large corpora. The human component of the research process is thus of paramount importance in this type of studies. The scope of the references consulted (e.g., Francis 1986; Hoey 1993; Flowerdew 2003; Mahlberg 2005; Lorés 2006), however, proves either too general or too specific. Francis (1986) and Flowerdew (2003), though looking at a wide range of nouns, are primarily concerned with the discourse and semantic aspects of these units. The data (presented as generalisations) tend to be unaccompanied by quantitative results. Besides, whilst the former looks at newspaper language, the latter examines a corpus of undergraduate biology lectures and textbooks. Mahlberg (2005) uses the BoE to describe the “local textual functions” (Mahlberg 2005: 59) of 20 general nouns. Such functions represent the meanings that general nouns acquire when used in context (e.g., *time* meaning ‘measurement’, ‘history’, ‘life’, etc.). Some general nouns qualify also as shell nouns (e.g., *thing, fact, system, problem*), whilst many others (e.g., *year, number, man, people*) are like shell nouns only on the basis of their unspecific meaning, but not in terms of their reference, inasmuch as only shell nouns may refer to stretches of discourse. Finally, Hoey (1993) and Lorés (2006), unlike the previous, provide more quantitative detail, but their findings are restricted to only the noun *reason* in Hoey (1993) and *thing*(s) and *idea*(s) in Lorés (2006).

Whilst showing that shell nouns represent a widely researched area, §2 has also cast light on a number of issues in need of further investigation. The literature reviewed calls for a study where i) all linguistic levels are considered (formal, syntactic, semantic and textual),
ii) a corpus-driven manual approach to data is foregrounded and iii) generalisations can be made based on a small but representative (and balanced) sample of the English language at large.

3. Methodology

This paper adopts a fully manual and corpus-driven methodology for the analysis of shell-noun data. Complete automaticity, as implied above, may prove beneficial in that it is quick and allows for large amounts of data from a large corpus to be feasibly processed. On the downside, however, the research scope may be considerably limited because of the restrictive nature of predefined automatic corpus queries, forcing the researcher to disregard certain linguistic features. In line with Sinclair (2004: 23), this study advocates an open approach to linguistic data, one which does not impose previous ideas on the analysis (e.g., shell nouns and noun complement clauses); one which, in short, “trust[s] the text”. Following also Hoey’s (2005) theory of lexical priming, shell nouns are not explored from just one linguistic perspective, as is often the case in the literature consulted. Instead, they are examined at all levels of linguistic analysis, i.e., formal/structural, syntactic, semantic and textual.

The study uses a small corpus of contemporary written and spoken British English, the BNC Sampler. Containing around 2 million words, the Sampler, a 2% sample of the entire BNC, is smaller, though considerably more balanced, than the BNC itself. This may be explained by the fact that written and spoken genres are more evenly distributed in the Sampler, with half of the words corresponding to each mode (50%-50%). The examples were retrieved from the web-based corpus analysis system CQPweb (Hardie 2012).

Subsequent to choosing the corpus, a list of 922 shell-like units was extracted from 15 lists available in the literature. Schmid’s

---

3 Winter (1977); Hoey (1979); Tadros (1985); Francis (1986); Ivanič (1991); Francis (1993); Winter (1992); Tadros (1994); Francis (1994); Francis et al.
(2000) 670-unit list is the most extensive list of shell nouns to date. Each of the 922 lemmas was sorted based on its overall token frequency in the BNC Sampler. The resulting frequency list was then divided into three main frequency ranges (i.e., high: 1, middle: 2 and low: 3), each containing 307 lemmas (except for 3, with 308 lemmas). Each group of units (high, middle and low frequency lemmas) was in turn subdivided into three further sub-ranges (i.e., upper: A, middle: B and lower: C), each, thereby, made up of 102 lemmas. The data analysed for this article comprise 1,110 concordance lines, those of which correspond to 42 shell-noun lemmas. For the sake of representativeness, a decision was made to choose lemmas from the three main frequency ranges (1, 2, 3) and then from the upper and middle sub-ranges within each main range (A and B). The present study relies on 13 units from the high range, 19 from the middle range and 10 from the low range. 10 units in each case belong to the middle frequency sub-range within each main range (i.e., 1B, 2B and 3B). 3 lemmas appear in the upper-high range (1A), 9 in the upper-middle range (2A), and none in the upper-lower range (3A). Table 1 below shows the 42 lemmas analysed and their distribution in the frequency list. As a general rule, a random selection of 20 concordance lines for the singular and 20 for the plural (40 in total) were analysed per lemma. This was not always possible, especially among low frequency lemmas, for which fewer than 40 hits appeared in the corpus.

---

(1998); Hunston/Francis (2000); Schmid (2000); Flowerdew (2003); Hinkel (2004); Flowerdew (2006).

4 As stated in §5, further research will provide a more even distribution of lemmas, with 60 in total (20 from 1, 2 and 3; 10 from 1A, 1B, 2A, 2B and 3A, 3B).
Table 1. Lemmas and frequency distribution.

For the analysis of each concordance line, the context of the node word was extended until a link could be established between shell noun and antecedent. More often than not, the intensive reading of long stretches of text was required in order to make sense of the reference of the noun. The difficulties encountered in delimiting antecedents underline the “fuzzy reference” (Francis 1994: 88) inherent in many cases of shell nouns, while also pointing to the apparent contradiction between the close reading of texts necessary for this kind of analysis and typical concordance-based corpus methods (Botley 2006: 102). Example (15) is a typical instance of the point just made. The shell-noun phrase the project does not refer to a specific stretch of discourse, but to all the information relevant to the understanding of its reference at this point in discourse:

(15)  […] and a Safety Authority which advises the Commission on all safety matters relating to the construction and operation of the Tunnel. In March the formal concession was signed Eurotunnel for the construction of two 7.6m diameter running tunnels and a 4.8m diameter service tunnel between Britain and France, and the operation of shuttle train service for the carriage of road vehicles. […] Neither the British nor French governments show any sign of bailing the company out although a future Labour government might take a different view. However, it seems unlikely that the international banking consortium will abandon the project given the large amounts of irretrievable money already committed. (BNC Sampler: BMJ, W:misc)
Each line was coded according to nine variables, reflecting the multi-faceted approach applied here. The following outlines the variables under scrutiny:

1. **Genre of the text**: The genre categories used in this study (e.g., W:ac, i.e., Written academic; S:meeting, i.e., Spoken meeting) are those taken from BNCweb (cf. Lee 2001).

2. **Experiential structure**: This variable concerns the semantic structure of the noun phrase, as explained by Systemic-Functional Grammar (Halliday/Matthiessen 2004). It comprises Deictic, postDeictic, Epithet, Classifier and Qualifier.

3. **Structural pattern**: The formal structure of the noun phrase is at issue here. The labels used for this variable (e.g., definite article, relative clause, noun complement clause, etc.) are adopted from Quirk et al.’s (1985) grammar.

4. **Syntactic function**: The framework followed here (e.g., direct object, subject, adverbial adjunct, object complement, etc.) is also Quirk et al.’s (1985).

5. **Participant type**: The analysis of the semantic roles played by nouns follows Halliday/Matthiessen’s (2004) system of Transitivity (e.g., Actor, Goal, Identifier, Carrier, Circumstance, etc.).

6. **Theme vs. Rheme**: This variable draws on Halliday/Matthiessen’s (2004) system of Theme.

7. **Reference**: This variable comprises the two main types of endophoric reference (anaphora and cataphora), their realisation (intersentential, intrasentential), as well as exophoric reference.

8. **Antecedent**: Following Stirling/Huddleston (2002: 1455), the term ‘antecedent’ is used here to refer to discourse encapsulated both anaphorically and cataphorically. The labels used in this variable are adopted from Gray (2010), where a distinction is made between Global Extended Discourse (where the antecedent is either difficult to delimit or crosses sentence boundaries) and Local Discourse, comprising Noun Phrase (simple: no postmodification and complex: with postmodification) and Sentence/Clause.
9. **Semantic type:** This variable concerns the meaning carried by a shell noun in context. The semantic classification followed here is the one proposed by Schmid (2000): factual, mental, linguistic, circumstantial, modal and eventive shell nouns.

4. **Findings and discussion**

As a starting point, five broad generalisations about shell-noun behaviour will be returned to here (see §2):

1. Academic and journalistic genres favour the use of this type of nouns.
2. Only specific deictics (e.g., the, this/that, these, etc.) frequently correlate with shell-noun phrases.
3. Structurally, NOUN + CLAUSE and NOUN + BE + CLAUSE are the two patterns that shell nouns tend to be associated with more often.
4. Anaphora lies at the centre of most research on these units.
5. Their antecedent is often realised by a stretch of discourse.

Each of these generalisations will now be compared with the findings from the present study. It is important to recall here the threefold distinction established above between fully automated, semi-automated and manual analyses (see §2.2), as reference will be made (whenever possible) to how each generalisation manifests itself across different methodological approaches.

4.1. **Genre: academic and journalistic genres?**

As far as genre is concerned, the sample reveals a predominance of written genres (66.39% vs. 33.60% spoken). A search for all nouns in the *BNC Sampler* shows a similar proportion (64.55% vs. 35.45%), proving that shell nouns share a similar mode distribution to nouns in
As regards spoken genres, three of these (i.e., S:meeting, S:conversation and S:speech) comprise 67.01% of the shell units employed in spoken English. In relation to written English, 51.15% of the examples in this mode is made up of three genres: W:newspaper, W:non academic and W:academic. Whilst the written distribution of shell nouns is similar to that of nouns (Biber et al. 1999: 65), their spoken distribution differs in that S:conversation and S:meeting contain a larger proportion of shell-noun use than they do with nouns in general. Figure 1 illustrates this point. The black line represents the spoken distribution of the 42 shell nouns in the study sample, while the grey line stands for the spoken distribution of nouns in the BNC.5

Figure 1. Distribution of nouns and shell nouns (Spoken).

As may be observed, only S:conversation and S:meeting peak above the line for general noun use in the BNC. Interestingly enough, the data obtained from conversation seem to contradict Biber et al.’s (1999: 65), insofar as nouns in that corpus-based grammar are claimed to be “by far least common in conversation”. This claim is also con-

---

5 The reason why the BNC Sampler was not chosen for comparison is that only Domain information (e.g., Business, Leisure, World affairs, etc.) is included in the corpus. Only BNCweb offers a genre classification of texts based on Lee’s (2001) categories.
firmed in the BNC, with conversation (of all genres) also showing the lowest percentage of noun use (12.51%). In the study sample, however, conversation appears among the top ten genres (7.30%) in a combined list of written and spoken genres. Similarly, S:meeting, with 8.47% of shell-noun use, features as one of the ten most frequent genres in the study sample, but as one of the least frequent (17.47%) as regards noun use in the BNC. A lexically-driven explanation might be advanced for such divergences, inasmuch as certain nouns may be more primed (Hoey 2005) for use in certain genres than others. One such noun is thing, whose “ubiquity […] in informal spoken conversation” relates to its presence in “such seemingly awkward expressions as the thing is that these children for instance are badly behaved ones usually” (Schmid 2000: 8). A similar example from the study sample is (16) below. The S:meeting texts analysed also seem to show a preference for some of the nouns under study, for example recommendation, project and application, which one would expect to occur frequently in business or institutional meetings.

(16)  <-|-> the thing <-|-> is in Scotland though, there's five good teams and the rest are rubbish. (BNC Sampler: KD6, S:conv)

4.2. The formal structure of shell-noun phrases: specific deictics and noun complement clauses?

Turning now to generalisations ii) and iii), the findings indicate that whilst a specific deictic (most commonly the) is highly frequent with shell-noun behaviour, non-specific deictics (most commonly a/an) and even the absence of a deictic may just as well be linked to this type of nouns. This is so much so that 65% of all determiners represent the definite article (27.02%), the zero article (22.11%) and the indefinite article (16.10%). Demonstratives are comparatively rare (5.74%), this being the most frequent alternative (2.83%, followed by that: 1.38%, these: 1.31% and those: 0.21%). Just as relevant is also the fact that it is not noun complement clauses but prepositional phrases that are in evidence in the sample examined. While prepositional phrases account for 59.83% of all post-head dependents in the data, noun complement
clauses comprise only 9.26% (to-infinitive clauses: 6.27, that-clauses: 2.99).

The high frequency of the indefinite and zero articles in the study sample seems to refute Schmid’s (2000: 25) claim that “indefinite noun phrases do not create as strong conceptual boundaries as the definite noun phrases in which shell nouns tend to occur”. Schmid’s (2000) full automation gives emphasis to a more corpus-based concordance-line analysis of examples, at the expense of a more intense reading of longer passages of discourse. Drawing on Ivanič (1991: 111), it is the latter kind of analysis that may bring to light such unexpected tendencies in the use of shell-noun phrases, insofar as even indefinite and uncountable noun phrases may function as shells provided that some context-dependent meaning is involved. This context-dependent meaning stems from what Garnham/Oakhill (1990: 380) refer to as “the mental model of the text”; in other words, all the information the reader or listener draws on as part of their attempt to interpret any referent appearing in discourse. The influence of context is evident in instances such as (17) and (18), where contextual saliency is involved despite the presence of an indefinite and uncountable shell-noun phrase respectively. The reference of a totally alien philosophy at the end of a speech in (17) may only be understood if one looks at the negatively loaded information presented in the preceding co-text. Similarly, in (18), the contention here is that some contextual information enables the reader to decode the nature of such a suspicion, in the sense that it is not suspicion as a general feeling that is at stake, but a specific suspicion that she is not what she looks like (i.e., not a servant, but a damsel sneaking away to meet her lover):

(17) But because they're reducing the fee income for dentists in the National Health Service these dentists, dedicated people, in some pretty tough estates in my patch, were saying, I am being forced to go private. I said, no, there's a development of a two tier health service. [...] Because what they're now doing to them is saying, unless you <pause> turnover two thousand prescriptions every month, we're gonna do away with your Professional Allowance. Well for the sake of a few million pounds the Conservatives will slash this area, that area, will say we need to close this unit, that unit. [...] We 're fighting a battle
against an enemy who has a totally alien philosophy to our to ourselves.

(BNC Sampler: H4A, S:meeting)

(18)  Outwardly, as Alianor saw her, she gave the appearance of a servant – but inwardly, to Joan herself, she was a damsel going to meet her lover. Letting herself out inconspicuously by way of the backstairs, Joan set out after dark for the Garden Tower, taking a circuitous route to avoid arousing suspicion.

(BNC Sampler: CCD, W:fict)

The dominant position that noun complement clauses occupy in the literature is not confirmed here, a finding also echoed by those references following more manual methodologies. Caldwell (2009: 176), for example, argues that of phrases are “by far the most prolific of all the patterns extracted as potential shell noun ‘hosts’”. This is also shown here, with 59.82% of all postmodifying prepositional phrases being headed by of. Similarly, Aktas/Cortés (2008: 10), in a semi-automated analysis of shell-noun use in professional and graduate research articles, extract “new patterns that had not been identified before”, some of which involve the preposition of (e.g., THE + NOUN + OF, A+ NOUN + OF). Flowerdew (2003: 337), more manual and qualitative in nature, posits that the lexical realisation of a shell noun may occur in the shape of nominal post-modification (see (19)).

(19) Nonetheless, you draw this distinction, despite the fact that the characteristic of both those areas is the characteristic of openness.

(BNC Sampler: FMP, S:pub_debate)

The prepositional phrase on its own, however, is not always informative enough, in which case its encapsulated information will have to be looked for elsewhere. Those instances where a prepositional phrase provides partial information about the reference of the shell noun offer only “specifics of identity” (Winter 1992: 154) (see (20)).

(20) Evil in Germany was so heinous, so unmistakable in the late 1930s, that all the world would assuredly join forces to wipe it out. Here was a nice example of the solution posing the problem.

(BNC Sampler: AEA, W:fict)
The prominence of prepositional phrases in the study sample is not specific to shell nouns. According to Biber et al. (1999: 606) “prepositional phrases make up 65-80% of all postmodifiers in all registers”, with *of* accounting for 60-65% of all prepositional phrases (Biber et al. 1999: 635). Featuring less frequently in the study sample are restrictive relative clauses (13.68%) *to*-infinitive noun complement clauses (6.27%), participle -*ed* clauses (3.28%) and *that* noun complement clauses (2.99%).

4.3. *Reference and antecedent: Anaphora and stretch of discourse?*

In relation to generalization iv), the findings reveal an overall dominance of cataphora in the study sample (42.70%), followed by anaphora (29.19%), exophora (24.14%), combinations of any of these reference types (3.42%) and unclear reference (0.54%). The salience of cataphora in the sample finds support in Sinclair’s (1993: 14) claim that discourse analysis ought to be primarily concerned with “forward-facing or prospective” reference. In his view, the processing of texts draws only on the current sentence, going so far as to claim that “[t]he whole text is present in each sentence” (Sinclair 1993: 14). Therefore, no benefit seems to be gained from looking backwards, as all the previous information should already be active in the readers’/listeners’ mind. It is only the information that the current sentence provides or is about to provide that the reader or listener is concerned with. Such a view is echoed by Winter (1977) and Tadros (1977, 1994), both focusing only on prospective or cataphoric shell-like units. While this approach is true in many cases of shell-noun behaviour, there are many other cases where, in line with the extensive literature on anaephora, the shell-noun is clearly retrospective, as in (17) above. Exophora (see (21) below), third in frequency, relates to those examples where no contextual information in the surrounding context enables a complete identification of the referent. Only if some background knowledge is available in the reader’s/listener’s mind will they be able to make some sense of the referent (Flowerdew 2003: 338; Ivanič 105-106). From the results obtained here, therefore, it may be argued that claims such as Francis’ (1994: 89) that “retrospective
labels are far commoner than advance labels” ought to be treated with caution, insofar as they may apply only to the corpus used (in this case, newspaper language).

(21) Since the 1980 Competition Act it is also possible for the Director General of Fair Trading to refer a particular practice of an individual firm to the Commission, where complaints have been received from supposedly injured parties, or the practice is suspected of limiting competition. In such cases a quick cost-benefit type appraisal and recommendation can be made. Where the practice is found to be anti-competitive the Minister can then accept voluntary undertakings from the firm to modify its practice or invoke extensive powers to proscribe the practice. (BNC Sampler: HXN, W:commerce)

Regarding now the link between noun and antecedent, the results show intrasentential cataphora as the most frequent subtype of reference (33.81%). This is in turn followed by intersentential anaphora (26.65%), intersentential cataphora (21.92%) and intrasentential anaphora (17.62%). Intrasentential reference is thus considerably more frequent with cataphora (see (22)), while intersentential reference is only slightly more frequent with anaphora (see (17) above). In simpler terms, this means that if a shell noun is prospective, we are more likely to find its antecedent within the same sentence. If it is retrospective, however, the chances are that its antecedent is going to be elsewhere. The literature reviewed does not offer any quantitative data in this respect. As stated above, however, the research focus is frequently placed on the intersentential functions of shell-like units.

(22) And our philosophy as a truly worldwide supplier of accounting software is that we should increase our presence in different parts of the world as our sales and marketing activity there merits it. (BNC Sampler: HDF, S:speech)

With respect to type of antecedent (i.e., generalisation v), the findings support the association between shell units and stretches of discourse. Using Gray’s (2010) taxonomy, 65.53% of all antecedents in the study sample are more complex than a phrase. This group of long antecedents is dominated by Global Extended Preceding Discourse
Formal, syntactic, semantic and textual features of English shell nouns

(31.06%) (see (17) above) and Clause (27.53%) (see (22) above). Sentential antecedents (see (23)) are comparatively rare (6.94%).

(23) China has shown some restraint in commenting on events in Eastern Europe and has made its views evident only in oblique ways. For example, the Chinese Communist Party pointedly sent a warm message of support to the conference of the hardline Romanian Communist Party. (BNC Sampler: A95, W:newsp)

It is important to remember here that global antecedents are cases where the antecedent spans two or more sentences, or cases where contextual information is present but some difficulty arises when trying to delimit the antecedent. Smaller antecedents are dominated by simple noun phrases (without postmodification) (see (24)) (13.66%), complements (8.48%) and complex noun phrases (with post-modification) (see (25)) (7.60%). The label ‘complement’, not present in Gray’s (2010) taxonomy, is used here to refer to appositive of phrases (see (19) above) and any other prepositional phrases providing full referential information (what the shell noun is) rather than specifics of identity (see (26)).

(24) Although rape and looting were technically crimes according to both armies rules of conduct. (BNC Sampler: H8W, W:essay)

(25) […] it does not make explicit provision for free and equal access to the media for all citizens – an equally desirable objective. (BNC Sampler: EBK, W:misc)

(26) The criteria for deciding whether the banks would support an application for funds led many banks to the conclusion that […] (BNC Sampler: G0C, W:commerce)

Only those scholars following a more manual methodology make some mention of smaller antecedents, most frequently noun phrases (e.g., Ivanič 1991: 11; Flowerdew 2003: 336). Gray’s (2010: 179) findings are similar to those obtained here, at least as far as non-phrasal antecedents are concerned (67% vs. 65.53% in the study sample). When looking at Gray’s (2010) four main antecedents, however, a different picture emerges, with clauses being slightly more frequent than global antecedents (35% vs. 32%), and complex noun phrases being considerably more common than simple noun phrases (26% vs.
7%). These results, nonetheless, cannot be compared with the findings in this study, because Gray’s (2010) scope is limited to sentence-initial this/these phrases in a corpus of Education and Sociology research articles.

4.4. Other variables

Of the nine variables explored in this study, three of them are not part of the five generalisations elaborated on so far, i.e., syntactic function, participant type and Theme vs. Rheme. The reason for this decision lies in the scant attention they receive in the literature reviewed. The results obtained for these three variables will now be summarized, drawing connections with the literature whenever possible.

Syntactically, the units examined show a marked preference for three functions, namely, direct object (see (26) above) (27.93%), subject (see (22) above) (20.18%) and subject complement (see (24) above) (10.81%). 70.81% of shell-noun occurrences in the sample perform a function at clause level, while only 29.19% perform a function at phrase level (i.e., complement of a preposition). Such findings, nevertheless, are not indicative of shell-noun behaviour, as they confirm Biber et al.’s (1999: 235) conclusion that the “relative frequency of nouns is much higher in object position and as a complement or object of a preposition than in subject position”. In line with Francis (1991: 147) and Hoey (1993: 82), it may be argued that, only by looking at individual units, will differences in syntactic function arise, as individual nouns will be more likely to appear in certain syntactic positions than in others.

Semantically, relational participants (in general) hold a dominant position in the study sample, comprising almost half of the examples (46.35%). They are, in turn, followed by circumstances (18.18%), material participants (12.96%), mental participants (7.65%), verbal participants (7.20%) and existential participants (5.04%). The clear dominance of relational processes (e.g., (21), (22), (24) above) in the present study stands in contrast with Matthiessen’s (1999: 14) findings based on a 14,500-word sample of written English. Though highly exploratory and tentative, the article provides some tendencies that
might reflect the frequency of Transitivity selections in a more representative sample of the language. His results differ from those obtained here in that they point to the dominance of material processes (51%), followed by relational (23%), verbal (10%), mental (9%) and existential (2%). From the evidence in this study, along with Vendler’s (1968: 73) connection between abstract nouns and the ‘NOUN is NOUN’ pattern (see (2) above), and Ivanič’s (1991: 97) recognition of “their role in the structure of relational process clauses”, shell nouns may be argued to prefer relational processes over any other process.

Finally, as regards the Theme-Rheme system, the analysis reveals a substantial majority of units occurring in Rheme position (e.g., (26) above) (75.05%), followed by Theme (e.g., (22) above) (19.55%), Marked Theme ((27) below) (2.88%) and Unclear (2.52%). The literature offers no evidence to either support or contradict this finding. Theme-Rheme is only referred to in passing in the analysis of particular examples and patterns. Flowerdew (2003: 333), for example, states that by placing a shell noun in Rheme position, the noun is highlighted as a “carrier of new information”. Hoey (1993: 77; 2005: 13), in turn, points out that Theme-Rheme distinctions are lexically specific, with certain nouns preferring clause-initial or clause-final positions.

(27) In this myth Isis was a woman who longed to increase her magic powers and join the gods in order to rule over them. She decided that the best means to this end was to learn Re’s secret name and thereby transfer his power to herself. […] (BNC Sampler: EVR, W:non_ac)

5. Conclusions

The results obtained in this paper suggest the importance of fully manual analyses of corpus data in certain areas of discourse analysis, especially those where the aim is to enrich our understanding of the overall linguistic and discourse behaviour of some units. This approach,
however, has certain limitations in that the close reading of long passages of discourse forces the researcher to reduce considerably the number of examples to analyse. A fully automated analysis has the advantage that it allows the processing of large amounts of data and is less taxing for the researcher, but at the same time it is less “revealing with regard to textual patterns” (Mahlberg 2005: 58). Drawing on Bednarek’s (2009: 21-22) three-pronged approach to corpus research, the present study would thus fall into “Small-scale corpus analysis”, in that a “context-sensitive analysis” takes precedence over the pattern-based automated analyses of lexical cohesion favoured in the literature on shell nouns. The corpus employed, however, is not as small as those found in this kind of studies (around 70,000-80,000 words), but one which, in Flowerdew’s (2011) terms, might be called a “large small corpus”, i.e., large enough to be representative of the language but small enough for the purposes of a manual approach to the analysis of lexical cohesion.

From the analysis conducted in this paper, a tentative and preliminary profile of shell-noun units may be presented:

1) **Genre**: This study confirms the prominent role that academic and journalistic English plays in shell-noun use. When compared to all other nouns, the evidence points to a similar generic distribution, with the only exception of S:meeting and S:conversation, which are more primed for shell-noun use than for general noun use.

2) **Deictic**: The results point to a preference for specific deictics, as also suggested in the literature. Indefinite and zero articles, however, appear to be just as important.

3) **Syntactic pattern**: Noun complement clauses, paramount in the literature, are only marginally frequent in the sample studied. Prepositional phrases clearly top the list of postmodifying structures, which proves unsurprising based on the behaviour of nouns.

4) **Participant**: There seems to be a strong association between shell nouns and relational processes, as revealed by the examples analysed.
5) **Syntactic function:** Direct object and subject are the two functions most frequently correlating with shell noun behaviour, a finding also typical of general noun use.

6) **Rhetorical function:** The considerable attention that anaphora receives in the literature does not appear to be confirmed by the findings here, showing cataphora to be the most frequent type of reference.

7) **Antecedent:** Though a clear relationship exists between shell-noun behaviour and long antecedents, the results here demonstrate that nominal antecedents may also appear.

8) **Theme vs. Rheme:** The vast majority of examples appear in Rheme position.

Further research will extend the scope of analysis from 42 to 60 lemmas, amounting to around 2,000 concordances. For the sake of an even distribution of units, 20 lemmas will be examined from each of the three main frequency ranges (see Table 1 above), i.e., top, middle and bottom. Different combinations of variables will then be explored (for example, Theme-Rheme and Structural Pattern), in search of any variations in shell-noun behaviour. Finally, the scope will be narrowed to the behaviour of individual lemmas across the three main frequency ranges, with a view to obtaining possible priming differences (Hoey 2005) among groups of shell nouns.

References


*BNC Sampler* = *British National Corpus Sampler*. <http://www.natcorp.ox.ac.uk/index.xml?ID=products#sampler>


From prototypical to peripheral: the ‘get + Ven’ construction in contemporary spoken British English

Abstract

The aim of the present study is to provide a preliminary approach to ‘get + Ven’ constructions in contemporary British English. As these constructions occur mainly in conversation, the study concentrates on the spoken part of the British Component of the International Corpus of English. It will, first, discuss the syntactic, semantic and pragmatic characteristics of central get-passives, as well as the similarities and differences between get- and be-passives; and, secondly, classify the get constructions identified in the corpus on a scale ranging from central get-passives to more peripheral constructions. It will examine the extent to which the distinctive features of get-passives apply to the rest of constructions with get and will also discuss the issue of which factors might explain the choice of different types of get constructions: presence of agent by-phrases, information status of subjects (animacy, responsibility), type of verb (activity, communication, mental, etc.), and the semantic nuance of the construction (beneficial (ben.), adversative (adv.), neutral (neut.)). My ongoing research suggests that differences between get constructions are especially noticeable with regard to the semantic prosody, the type of verb and the degree of responsibility of the subject.

1 Author’s email address: eduardo.coto@usc.es; affiliation: U. of Santiago de Compostela. For generous financial support I am grateful to the European Regional Development Fund and the following institutions: the Spanish Ministry of Economy and Competitiveness (grants HUM2007-60706 and FFI2011-26693-C02-01), the Autonomous Government of Galicia (grant CN2012/012), and the Spanish Ministry of Education (FPU grant 2009-3554). Thanks are also due to two anonymous reviewers and to Teresa Fanego, Elena Seoane and Paloma Núñez for their feedback on an earlier version of this article, as well as to the audience of my presentation at the CILC2012 conference, who provided me with new ideas that improved the final version of this article.
1. Introduction

The present study is part of a larger project consisting of a cross-varietal comparative analysis of ‘get + past participle’ constructions in East and Southeast Asian Englishes. In this study, the focus is on get constructions in contemporary spoken British English, on the basis of data extracted from the British Component of the International Corpus of English (ICE-GB). In §2, I will classify the different get constructions on a gradient according to their degree of passiveness, which ranges from more to less prototypical. Section 3 examines the distinctive features of the most prototypical subcategory, namely central get-passives, in order to check the extent to which these characteristics apply to the rest of subcategories. Section 4 describes the corpus used in this study and presents the results of the analysis. A brief summary of the main conclusions is presented in §5. Given the compatibility across the individual corpora in ICE, I will use the findings obtained in this study as a benchmark study for further research, when comparing British English with other varieties, such as Indian English, Hong Kong English, and Singaporean English.

2. The passive gradience: from prototypical to peripheral

A number of linguists, among them Granger (1983: 103), Quirk et al. (1985: 167-171), Svartvik (1985: 138) and Collins (1996: 45), claim that the definition of the English get passive, that is, the construction ‘get + past participle’ or ‘Ven’, is very broad and that the various get constructions form a fuzzy set. The get-passive has been regarded as a “linguistic puzzle” (Carter/McCarthy 1999: 54), as “a contentious point of discussion” (Chappell 1980: 411), or even as “the subject of widespread disagreement” (Collins 1996: 43). In fact, the term passive is misleading, since there is a heterogeneous number of ‘get + past participle’ constructions that involve types of meanings different from
the regular passive functions, despite their formal and semantic similarity. In order to classify these constructions, the above authors place them on a gradient according to their degree of passiveness. Although there is no agreement on an exact gradient, the following hierarchy, based on the work of these authors, seems to be fairly comprehensive.

2.1. **Central get-passives**

At the top end we find the structures that Quirk et al. (1985: 167) call “true passives” and Collins (1996: 45) “central”. The latter term will be used in the current study. According to Collins (1996: 49), this “central” subtype is the most strongly represented of the six subclasses, whose constituents have a direct active-passive relation, as in example (1):

(1a) Uncle Ahmed **got bitten** by the snake. <ICE-GB:S2A-047 #52:1:A>
(1b) The snake bit uncle Ahmed.

Collins (1996: 45) proposes a list of features which – he argues – are definitory of central get-passives. These are the presence or absence of an agent phrase, the semantic properties of the agent (animate vs. inanimate; human vs. non-human), the accessibility (if not present) of the agent, and the stative/dynamic meaning of the lexical verb. Get-passives with a human agent rank highest, as in *He got invited to the party by the host,*\(^2\) since the association between agency and intentionality is stronger among human than among non-human agents, as in *The car crash got caused by a reckless deer.* Inanimate entities, as with *lightning* in *The poor man got struck by lightning,* are not associated with intentionality at all, though the active-passive alternation is still possible: *Lightning struck the poor man.* Ranked below these are constructions which have a preposition other than *by* (cf. Palmer 1988: 87-88; Collins 1996: 46), notably *about, at, over, through, to or with.* In an example like *He got hit in the nose with a golf ball,* the prepositional phrase following the get-passive does not

\(^2\) Throughout the chapter, examples whose source is not explicitly mentioned are made-up examples.
resemble an agentive *by*-phrase because it takes a different preposition, but semantically it is somehow similar. This passive clause could have two active counterparts, as shown below; example (2a) would correspond to an agentive interpretation of the *with*-phrase, while (2b) would involve an instrumental reading of the *with*-phrase:

(2a) A golf ball hit him in the nose.
(2b) Someone hit him in the nose **with a golf ball**.

Prepositional phrases like these, with a twofold equivalent in active counterparts, are what Svartvik (1985: 104-105) calls “Janus-agents”. He defines them as follows:

Some prepositional phrases may have twofold potential function in active transforms: their nominal part may function as active clause subject, or the entire prepositional phrase may function as adjunct (usually instrumental, ‘by means of’) with some other nominal element as active clause subject. Such adjuncts which permit two different active clause transforms [...] will be called ‘JANUS AGENTS’. (Svartvik 1985: 104-105)

Although the possibility of an active counterpart suggests that there is a fine line between agentivity and instrumentality, agent phrases introduced by a preposition other than *by* (as opposed to agent *by*-phrases) are not related at all to volition or intentionality (of a generally animate and human agent), and thus they are not agentive, but rather instrumental. Hence, I will not analyse agent-like prepositional phrases as agent phrases here, since strictly speaking they are not such.

Further down the scale are the agentless *get*-passives, which have no expressed agent, as in the following example:

(3) They said you don’t take any pictures otherwise you you [sic] **get killed** And there were threats all round <ICE-GB:S2A-050 #148:2:A>

An overwhelming majority of *get*-passives occur without an agent, and a possible explanation for this is that, since some measure of responsibility for initiating the action is ascribed to the subject-referent in *get*-passives, the role of the subject-referent in the process is of greater concern than that of the agent, which might have low infor-
mation value, for instance, when it is obvious, pragmatically inferable, unknown or irrelevant (cf. also Palmer 1988: 78-79). These factors lead Collins (1996: 46) to specify a scale, going from cases where the agent, though unexpressed, is salient in the linguistic context, as in *The young girl got hit by a truck and got killed* where the agent-phrase *by a truck* is readily reconstructible, or in *The carpet’s loose there and my heel got caught* (from Collins 1996: 45) where *by the (loose) carpet* is easily inferable from the linguistic context, to cases where the agent is merely implicit and might be figured out without complete certainty, as in *We cut the amount of wood the boss required, without getting caught cutting without a licence* (from Collins 1996: 45), in which we might infer, in part from the rest of the sentence and in part from general knowledge, that the indefinite and non-specific agent is human and has some authority, for example, the police.

2.2. **Semi get-constructions**

The next type of get constructions, also known as “semi-passives” (cf. Quirk et al. 1985: 168) and “psychological get-passives” (cf. Collins 1996: 46), belong to a subclass whose members show both verbal and adjectival properties:

(4) I [sic] I [sic] always get really turned on to the the [sic] style of it by the look of it <ICE-GB:S1B-023 #56:1:B>

Example (4), for instance, is verb-like in having an agent by-phrase and an active analogue (*The look of it always really turned me on to the style of it*). On the other hand, the past participle shows several adjectival properties including the possibility of: i) modifying the participle with an intensifier such as *extremely, quite, rather or very* (*really turned on*); ii) coordinating the participle with an adjective (*I got turned on by and anxious about...*); and iii) replacing get with a lexical copular verb such as *become, feel or seem* (*I became turned on by...*). It is worth mentioning that most of the participles in this subclass are stative rather than dynamic, which favours an adjectival analysis, since all participial adjectives have a stative meaning,
whereas corresponding verbs are usually dynamic. Quirk et al. (1985: 168-169) and Collins (1996: 46) remark that the use of agent by-phrases in these adjectival past participle constructions is very rare, although agent-like phrases introduced by different prepositions seem to be quite common, among them about, at, over, through, to and with, as in So you got involved through her <ICE-GB:S1A-081 #121:1:A>. As mentioned above, when dealing with ‘Janus agents’ in central get-passives, these prepositional phrases are not strictly speaking agent phrases and hence are not representative of the passive voice, and for this reason the term agent-like phrase is an appropriate one.

2.3. *Pseudo* get-constructions

Further down the scale are pseudo get-constructions (also called “pseudo-passives”, cf. Quirk et al. 1985: 169-170; and “reflexive” or “reciprocal passives”, cf. Collins 1996: 47), which seldom have an active counterpart and in which the possibility of agent addition is very rare, as in *I have to get dressed before eight o’clock* (from Quirk et al. 1985: 161) or *[I]f you ever got married I’d be absolutely devastated <,> <ICE-GB:S1A-050 #156:1:B> (other examples include changed, finished, shaved, started, washed, etc.). In these constructions, get functions as a copular verb and not as a passive auxiliary, and is followed by a stative past participle (e.g., dressed and married in the examples above, according to Quirk et al. 1985: 161 and Collins 1996: 47, mean ‘in a state of marriage’ and ‘in a state of wearing clothes’, respectively) as its complement, which is therefore adjectival. According to Palmer (1988: 88), “[t]he -en forms that function in this way are essentially perfect in meaning and refer to a resultant present state […].” Nonetheless, these participles do not fulfil all of the properties attributed to adjectives: for instance, they are not gradable (*They got very washed*), get cannot be replaced by a lexical copular verb (*They seemed washed, although ?They seemed married*), and the criterion of coordinating the participle with an adjective is unclear (?They got washed and ready). Finally, as in get-passives,
get puts the emphasis on the subject and on what happens to him or her as a result of the action.

2.4. Adjectival get-constructions

Close to the periphery we find adjectival get-constructions (termed “adjectival passives” by Collins 1996: 48) whose members are adjectives, derived by lexical-morphological conversion of the Ven form, and get is a copula rather than a passive auxiliary. These constructions commonly involve a change of state rather than an agent-motivated event, which is why the presence of an agent by-phrase is usually impossible. As with central get-passives, this subclass is not homogeneous, but represents a scale of degree of adjectivalisation. The more criteria a given example fulfils, the closer it is to a prototypical adjective, namely the possibility of being used attributively, premodified by a degree adverb and coordinated with another adjective, or of replacing get with a lexical copular verb:

(5) She got entangled with this book.

Example (5) fulfils almost all the above-mentioned criteria: premodification (She got very entangled), coordination with another adjective (She got entangled and pleased), replacement of get with become (She became entangled), though it cannot be used attributively (*An entangled book). Participles like bored, drunk, frightened and tired, which refer to human states, move yet further along the scale from verbal to adjectival and are close to prototypical adjectives in that they enjoy a greater freedom of occurrence than the participles discussed above with premodification (I mean I get really tired; <ICE-GB:S1A-018 #329:1:A>) and in predicative use (A frightened old man).
2.5. *Idiomatic* get-constructions

On the very periphery of the get-constructions gradient we find idiomatic expressions where the relationship between the Ven form and the verb from which it historically derives has been totally lost from sight, as with *get stuck into* in *There’s a Jewish tradition* <,> of uh [sic] scholarship in which after a certain stage you just have to get stuck in to public life <,> <ICE-GB:S1B-047 #14:1:B>. Similar idiomatic expressions include *get accustomed to*, *get fed up with*, *get rid of* and *get used to* (cf. Collins 1996: 49; Leech et al. 2009: 156).

2.6. *Reflexive* get-constructions


(6) Isn’t it about time that the Department really *got itself organised* to avoid all these unnecessary deaths which are taking place which could be matched providing there was a real determination by the minister to do a proper recruitment policy <ICE-GB:S1B-055 #9:1:C>

These constructions, illustrated by example (6) above, which have *get* as a main verb taking non-finite complementation, cannot be considered as *get*-passives, since the passive *get* construction is a simple catenative with no intervening NP between *get* and the non-finite complement (cf. Huddleston/Pullum et al. 2002: 1443). When the intervening NP is a reflexive, the construction is explicitly agentive, with the subject-referent being involved very directly and being responsible for the action described. According to Chappell (1980: 430-431), the adversative reflexive is more appropriately used, depicting a subject as “tempting fate through provocative actions or perhaps what the speaker judges to be some kind of ill-considered behaviour,
expected to lead to an unfortunate state of affairs”, as in *John gets very reckless when he’s drunk, that’s why he got himself run over*. In beneficial reflexives some kind of intentional causation on the part of the subject is clearly visible, as in *She got herself promoted*, through her dedication or hard work, for instance. According to Sussex (1982: 86) and contrary to Chappell’s (1980) analysis, reflexive constructions can even occur with inanimate and abstract nouns, as in *My theory got itself accepted at the annual convention*. The difference between the reflexive and the plain *get*-passive construction is that the former suggests some kind of involvement which is more than pure passive suffering of an action, but less than agentive involvement. There are occasions on which the intervening NP is not a reflexive, as in *My sister and I were going to get a picture of of [sic] she and I done* <ICE-GB:S1A-015 #48:1:B> or *One commentator getting his his [sic] metaphors slightly mixed wrote about the the [sic] virtual canonization of Lord Jakobovits by the Prime Minister <,> <ICE-GB:S1B-047 #41:1:A>. Again, these constructions, where *get* functions as a causative verb followed by an object and an -ed participle complementation, cannot be considered passives.

Summarising, we can distinguish six types of constructions with *get*: first, central *get*-passives, with the auxiliary *get* followed by a verbal past participle; second, semi *get*-constructions, which show both verbal and adjectival properties; third, pseudo *get*-constructions, which have neither an active counterpart nor an agent phrase and are followed by a stative past participle; fourth, adjectival *get*-constructions, whose members exhibit adjectival properties; fifth, idiomatic *get*-constructions; and finally reflexive *get*-constructions, with an intervening NP between *get* and the past participle.

3. Characteristics of central *get*-passives

Central *get*-passives are best described by comparing them with the most widespread passive periphrasis: the *be*-passive. As opposed to
be-passives, get-passives tend to be avoided in formal English, yet are typically recurrent in conversation (cf. Quirk et al. 1985: 161; Biber et al. 1999: 476; Huddleston/Pullum et al. 2002: 1442; McEnery et al. 2006: 112-113), do not normally have an overt agent phrase (cf. Quirk et al. 1985: 161; Carter/McCarthy 1999: 52; McEnery et al. 2006: 113), and occur only with dynamic verbs (cf. Huddleston/Pullum et al. 2002: 1442; Alexiadou 2005: 17). Moreover, the animate subject of the get-passive is usually responsible for initiating the event and is also commonly attributed adversative consequences in these constructions (cf. Hatcher 1949: 436-437; Collins 1996: 52; Carter/McCarthy 1999: 49-50). There follows a close examination of some of these characteristics.

3.1. Presence of an agent by-phrase

As the passive voice is often used as a strategy that allows speakers to avoid mentioning the agent, it might be expected that agentless passives are significantly more common than those with an overt agent. Quantitative considerations are a supporting argument here; for instance, Jespersen (1933: 121) mentions that “over 70 per cent of passive sentences found in English literature contain no mention of the active subject”. On the same lines, Quirk et al. (1985: 164) state that “approximately four out of five English passive sentences have no expressed agent”, and that get is usually “limited to constructions without an expressed animate agent” (Quirk et al. 1985: 161), as in James got beaten last night. The findings of McEnery et al.’s (2006: 113) study show that in the Freiburg-LOB Corpus of British English (F-LOB) (Mair/Leech (2007), of approximately one million words, the short form of the be-passive is over eight times as frequent as its long form, while for the get-passive the short form is over ten times as frequent as the long form. The contrast in the spoken part of the British National Corpus (BNC), of about 10 million words (10% of the corpus), is even more marked, where the short forms of be- and get-passives are over 18 and 37 times as frequent respectively as their long forms. The data show, then, that short passives are clearly more common than long passives in both spoken and written English, that
short passives are also significantly more frequent in spoken than in written English, and that get-passives are more likely than be-passives to occur without an agent.

Furthermore, the information by-phrases normally provide is new rather than given (or old) and for that reason such phrases are a key element in the discourse. According to Givón (1990: 897), “by ‘old’ one means ‘assumed by the speaker to be accessible to the hearer’ and by ‘new’ ‘assumed by the speaker not to be accessible to the hearer’”. The very low frequency of get-passives with an expressed agent by-phrase may also be related to the type of subject. Givón (1993: 69) holds that “while the subjects of be-passives divide roughly equally between human and non-human ones, the get-passive shows a lopsided 9:1 preference for human subjects”. He relates this to the “agentive nature of the patient-subject of the get-passives” (Givón 1993: 68).

3.2. Type of lexical verb

Another difference between be-passives and get-passives involves the type of lexical verb they contain. Huddleston/Pullum et al. (2002: 1442) state that get-passives are found only with dynamic verbs, that is, those verbs which denote an action and not its outcome, while be-passives can be found both with dynamic and stative verbs. In the same way, Alexiadou (2005: 17) argues that “the get-passive is not permitted with stative verbs and verbs that do not allow for the subject of the construction to be interpreted as affected”. Be cannot, then, be replaced by get in an example like It was/*got believed that the letter was a forgery (from Huddleston/Pullum et al. 2002: 1442).

3.3. Responsibility and animacy of the subject

The subject in central get-passives is usually attributed some kind of responsibility for initiating the event described in the clause, which is determined both by the meaning of get and by features of the context. As Huddleston (1984: 445) has noted, “get lends itself more readily
than *be* to the imputation to the subject-referent of some measure of initiative or responsibility”. While a clause with *get* would be uttered by someone directly involved or affected by the action described, the equivalent clause with *be* would be more appropriate for a person trying to present an objective account of the event. Example (7b) below shows that the subject of the *get*-passive is in some way responsible for his own misfortune, while no such inference can be drawn about the subject of the *be*-passive in (7a):

(7a) He was caught by the police.
(7b) He got caught by the police.
(from Givón/Yang 1994: 120)

The extent to which the subject is in control of the event described differs, creating a scale in the construction which is best exemplified by Vanrespaille’s (1991: 107) *agentivity gradience*. The subject is marked on a gradience of agentivity which goes from the least agentive (*He got arrested*), passing through an intermediate agentive (*He got worried*), to the most agentive (*He got dressed*). Vanrespaille (1991: 104) describes this range as from “a mere hint of responsibility with a human subject over reflexive activity to causation on the part of the subject”.

Since it is animate human referents that can show volition and intentionality over their actions (cf. Dahl/Fraurud 1996: 58), the concept of subject animacy is clearly related to the notions of subject control and responsibility. *Get*-passives are eventive in terms of their aspect and this fact presumably contributes to the animacy effect; since events are usually controlled by an actor, animates are more likely to be able to control those events. Thus, if speakers talk about animate beings, there is a greater chance for them to use the *get*-passive rather than the *be*-passive. Evidence for this is shown by Arce-Arenales et al. (1994: 14), who report that of the 32 *get*-constructions with a passive interpretation in their study, 88% have human subjects, compared to 47% for the *be*-passives. On similar lines, Toyota’s (2007: 153) findings on the animacy of the subject entity in *get*-passives show that only 14.3% of the subjects are inanimate while 85.7% are animate, of which 84.7% are human. To a lesser extent, and contrary to Lakoff’s
(1971: 154-155) assertion that they cannot take inanimate subjects, get-passives can indeed take inanimate subjects. Givón/Yang (1994: 121) explain that

when the subject of the get-passive is inanimate, thus itself incapable of responsibility, some human associated with the subject, or with the event in some capacity, may either retain responsibility, be involved in the action or be adversely affected by the results.

The affected entity in these constructions is not the inanimate subject, but rather the person who owns or is responsible for it, as in the case of Martin in example (8) below. However, there are cases in which the non-human subject-referent does not stand in a direct relationship to any person or persons, as in example (9), and in this case neither is the natural phenomenon of the rise of sea levels subject to human control:

(8) Martin’s video recorder got fixed last week.
(9) And because the sea levels rise a lot of places could get flooded Mm [sic]. <BNC FMR S_classroom>

3.4. Attribution of beneficial and adversative consequences

This implication was first noted by Hatcher (1949: 441), who claimed that central get-passives were only used for two types of events, “those felt as having either fortunate or unfortunate consequences for the subject”. This dichotomy is illustrated in the following examples from Collins (1996: 52), from the Brown corpus (Francis/Kučera 1964); example (10) is an instance of an adversative connotation, in that getting fired is unfavourable to the subject affected by the event, while (11) is an example of a beneficial implication, since getting promoted is favourable:

(10) ‘We got fired’, Jones said. <Brown-N01-1650>
(11) Some of them were warts until they got promoted. <Brown-M01-1440>

The majority of get-passives refer to adversative contexts, “a state of affairs that is signalled contextually by the conversational participants as unfortunate, undesirable, or at least problematic” (Carter/McCarthy
1999: 49). Furthermore, this adversative/beneficial distinction can also be applied to *get*-passives with non-human subjects. Chappell (1980: 440) argues that “the affected entity in this construction is not the inanimate subject, but rather the person who owns it, or else stands in a relationship to this object equivalent to that of ownership”, as in *Jane’s bike got stolen*.

4. A corpus-based study of *get*-constructions

4.1. The corpus and the database

Having presented the classification of *get* constructions and the characteristics of each of the subtypes, I will now turn to the empirical part of my study, which involves the identification and analysis of the different constructions with *get* in a corpus of British English, in light of the characteristics described above. As these constructions are found mainly in conversation, I concentrated on the spoken part of the British Component of the International Corpus of English (*ICE-GB*). *ICE-GB*, as with the rest of the individual corpora in *ICE*, contains 500 samples (both spoken and written English) of approximately 2,000 words each, leading to a corpus of around one million words. The spoken component (300 samples) consists of dialogues (180), both private (100) and public (80), and monologues (120), both unscripted (70) and scripted (50) (files S1A, S1B, S2A and S2B, respectively), amounting to around 600,000 words. The texts in the corpus date from between 1990 and 1993 and the authors and speakers (both male and female) are aged 18 or over and were educated in English. *ICE-GB* in particular is both tagged for word class and parsed syntactically and comes with the ICECUP III retrieval software, which is very useful and practical for complex and detailed searches across the whole corpus. First of all, I carried out an automatic search with the ICECUP software for the different word forms of *get*, namely *get*, *gets*, *getting* and *got*. A manual search was then carried out for these forms followed by a past participle. All the relevant examples retrieved were
entered into a database, then classified following the passive gradient above and analysed according to a number of variables.

4.2. The results of the analysis

The quantitative analysis of the ICE-GB corpus yielded a total of 215 tokens of *get* (get=114, gets=59, getting=24, got=18) followed by a past participle, whose distribution according to type of construction is shown in Table 1 below:

<table>
<thead>
<tr>
<th>TYPE OF CONSTRUCTION</th>
<th>PERCENTAGE</th>
<th>TOKENS</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>get</em>-passives</td>
<td>23.26</td>
<td>50</td>
</tr>
<tr>
<td>semi <em>get</em>-constructions</td>
<td>6.05</td>
<td>13</td>
</tr>
<tr>
<td>pseudo <em>get</em>-constructions</td>
<td>13.49</td>
<td>29</td>
</tr>
<tr>
<td>adjectival <em>get</em>-constructions</td>
<td>20.93</td>
<td>45</td>
</tr>
<tr>
<td>idiomatic <em>get</em>-constructions</td>
<td>16.74</td>
<td>36</td>
</tr>
<tr>
<td>reflexive <em>get</em>-constructions</td>
<td>19.53</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>215</td>
</tr>
</tbody>
</table>

Table 1. Distribution of *get* + *Ven* according to type of construction.

As can be seen, although all constructions occur in reasonably large numbers in the corpus sample, the *get*-passive is the most commonly represented of the six subclasses, closely followed by adjectival and reflexive *get*-constructions. I illustrate the gradience from prototypical to peripheral with two examples from each of the six subclasses:

**Get-passives**

(12a) And if you’re going to have them made of mud brick and it rains remember to actually stick a ledge or cornice all the way round <> so that the <> mud can not actually get ruined by the rain <ICE-GB:S2A-024 #80:1:A>

(12b) I can’t see much point in doing it at the moment because <> you don’t get *paid* for it <> <laugh> <ICE-GB:S1A-066 #139:1:B>
Semi get-constructions

(13a) Well I got involved through Celeste who <> I used to dance with with [sic] London Contemporary <> going back <> twenty years or so <> <ICE-GB:S1A-002 #4:1:B>
(13b) So after we actually did Islington it it [sic] sort of fell in Canonbury it fell into place and they got quite excited about the Elizabethan side of it <> <ICE-GB:S1A-094 #12:1:A>

Pseudo get-constructions

(14a) I suppose we had to get engaged now because it’s the only way we’re going to see each other <ICE-GB:S1A-050 #182:1:B>
(14b) I’m I’m [sic] def [sic] I’m [sic] definitely not going down that road today because we’ve got to get finished by Christmas <ICE-GB:S1B-029 #111:1:A>

Adjectival get-constructions

(15a) Yeah but the thing with Trivial Pursuits I get bored after a while <ICE-GB:S1A-048 #163:1:A>
(15b) You must be very careful with that cos [sic] otherwise you’re going to get confused <ICE-GB:S1B-015 #192:1:A>

Idiomatic get-constructions

(16a) First of all you don’t score much and secondly you only get rid of two letters <ICE-GB:S1A-010 #162:1:B>
(16b) As you’re you’re [sic] getting used to it you’re getting better aren’t you <ICE-GB:S1B-004 #82:1:A>

Reflexive get-constructions

(17a) Far from being cold-hearted the great eighteenth century rationalist republican Thomas Paine who’s been called the greatest Englishman got himself sentenced to the guillotine because he argued passionately in the assembly of the French Revolutionary Council for the life of King Louis <> <ICE-GB:S2B-032 #15:1:A>
(17b) Well he’s got his fir [sic] first poster done and stuff like that <,> <ICE-GB:S1A-069#277:2:B>
4.2.1. The agent by-phrase

Although the prototypical get-passive is accompanied by an explicit agent by-phrase, the corpus contains just eight tokens of agent by-phrases, three of these occurring in get-passives, three in reflexive get-constructions and two in semi get-constructions, as the examples below illustrate (example (1) is repeated here as (18) and example (4) as (20), for convenience). The term prototypical should be understood here in the sense of ‘archetypical’ or ‘ideal’, and not as ‘most common’ or ‘habitual’. That is, a get-passive taking an overt agent by-phrase is the ideal passive construction in that it fulfills all the defining criteria for central get-passives, but is not obligatorily the most frequent construction. In fact, an overwhelming majority of get-passives tend to leave the agent phrase unexpressed. This confirms the suggestion of Quirk et al. (1985: 161) and also Carter/McCarthy (1999: 52) that get-passives (as well as the other subtypes) are generally agentless, mainly because of the low information value the agent has.

Get-passive

(18) I know my period started the year that Uncle Ahmed got bitten by the snake but it doesn’t help you to actually identify that particular year <,> <ICE-GB:S2A-047 #52:1:A>

Reflexive get-construction

(19) Then I came back and on the way back I got a Falange pass found on me by the Muslims who wanted to cut my throat <,> <ICE-GB:S2A-050 #142:2:A>

Semi get-construction

(20) I [sic] I [sic] always get really turned on to the the [sic] style of it by the look of it <ICE-GB:S1B-023 #56:1:B>

The long passive (where the agent phrase is expressed overtly) retains all the information that would be expressed in its corresponding active counterpart. The function of long passives, therefore, is not that of eliding information, but that of rearranging the order of constituents in
the clause, mainly for reasons of the length and information status of such constituents (cf. Seoane 2009). For this purpose, I also examined the length of the subject and the agent by-phrase in these eight agentful constructions and found that the choice of the long passive can to a large extent be accounted for by the principle of end-weight, according to which long and heavy elements tend to be placed at the end of the clause, as in examples (19) and (20) above. In fact, in seven of the eight agentful constructions the agent phrase is longer than the subject in terms of number of words (note that the preposition by has been counted as part of the agent phrase). However, there is a balance as regards the information status of these constituents, since both subjects and agents tend to provide given information. The general tendency, nevertheless, is for subjects to have a higher degree of givenness than the agent by-phrase, which usually convey new information, retaining the unmarked given-before-new order of constituents in clause structure. Likewise, as far as the animacy of the agents is concerned, the proportion of animate and inanimate agents is the same, which contrasts with Dahl/Fraurud’s (1996: 58) expectations that the referent of the agent by-phrase be animate, given the strong association between agency and animacy. Nonetheless, there is a preference for human (three tokens) over non-human animate agents (one token) (example (18) above).

4.2.2. The dynamicity of the verb category

One factor which might determine or favour the use of one particular get construction or another is the semantic type of verb used. For this reason, it is important to classify the verbs in my data from a semantic point of view. I have followed Biber et al.’s (1999: 360-364) classification, which distinguishes seven major semantic domains: activity verbs (buy, put, send), communication verbs (ask, tell, write), mental verbs (know, see, think), verbs of facilitation or causation (allow, cause, enable), verbs of simple occurrence (become, happen, occur), verbs of existence or relationship (appear, exist, seem), and aspectual verbs (continue, finish, start). This classification is based on the core meaning of the verb, that is, on the meaning that speakers first associate with a given verb. An analysis of the verb type distribution in my
corpus sample according to this classification showed that the overwhelming majority (70%) of the verbs occurring with get constructions are activity verbs (see Figure 1 below). It is worth noting that 46 of the 151 activity verbs in the corpus occur in get-passives, while 21 of the 53 mental verbs appear in adjectival get-constructions. This overwhelming frequency of activity verbs was expected since, in general, the category of activity verbs occurs much more frequently than any other verb category and they are particularly common in conversation (cf. Biber et al. 1999: 365-366). In addition, get-passives, as opposed to be-passives, tend to occur with activity verbs, which are the dynamic verbs par excellence (cf. Huddleston/ Pullum et al. 2002: 1442). On the contrary, mental verbs yield stative participles or adjectival -ed forms, which favours an adjectival analysis. The semantic categories of communication and aspectual verbs were recorded to a lesser extent, while not a single instance of simple occurrence, causative or existence verbs was found. This is not surprising, since causative verbs are relatively rare and get is not possible with verbs which report a state of existence or a relationship that exists between entities, such as appear or seem, exist or live, as in *He got seemed to be late.

Figure 1. Distribution of get + Ven according to verb type.

4.2.3. The adversative semantic nuance

An examination of the semantic nuance conveyed by get constructions in the corpus sample yielded the following findings: 65 examples
(30%) of adversative connotations, 48 examples (22%) of beneficial implications, and 102 examples (48%) of neutral value for the subject-referent. What is meant by “neutral value” is that there are no beneficial or adversative effects upon the subject, or, at least, those effects are not clearly visible, as in get gathered, get printed out, get sent, get sprayed, get told and get transcribed. Of interest here is the preference for adversative (30%) rather than beneficial (22%) implications for the subject (cf. also Carter/McCarthy 1999: 49). This evident preference for unfortunate consequences is confirmed not only for get-passives (8 ben./22 adv.), as in get caught, get hurt, get inundated, get killed, get ruined and get struck, but also for adjectival get-constructions (5 ben./30 adv.), as in get annoyed, get bored, get confused, get frustrated, get scared and get tired. Nonetheless, the balance is tipped in favour of beneficial implications in the case of reflexive get-constructions (22 ben./5 adv.), as in get it organised, get it sorted out, get that saved, get the job done and get their money paid (see Figure 2 below). Although not as significant as with reflexive get-constructions, semi and idiomatic get-constructions occur slightly more frequently with beneficial than with adversative connotations (5 ben./4 adv. and 7 ben./3 adv., respectively). Pseudo get-constructions showed a balanced use of favourable and unfavourable situations (1 ben./1 adv.) with a well-defined predominance of neutral utterances (27 out of 29). As already mentioned, reflexive get-constructions, despite their name, are very much on the periphery of the get-passive category. As shown in Figure 2, this class is also semantically different from the rest of the constructions.
4.2.4. The subject: degree of responsibility and animacy features

Two aspects of the subject in get constructions have been analysed: first, whether it can be attributed some degree of responsibility for the action described in the clause, and, second, its animacy features. As regards the former, the results confirm that in almost 67% of the occurrences analysed the subject is definitely responsible for initiating the action (example (21) below), whereas only in 27% of the cases is the subject clearly not in control (e.g., (22)). Cases in which it is not clear whether the subject is in control and in which the context is ambiguous were classified as neutral (6%), as in example (23):

(21) Everybody knows [that] some politicians get drunk some politicians <,> uh [sic] uhm [sic] sleep around <,> <ICE-GB:S1B-024 #125:1:F>

(22) It escaped on the underground and it got out this poor wasp so far <,> far from home And it was on the window and I thought pe [sic] perhaps I ought to kill it because somebody’ll [sic] get stung and then I thought no why shouldn’t it have at least a chance even if it does sting everybody So I I [sic] just left it <ICE-GB:S1A-067 #49:1:A>

(23) London London [sic] uhm [sic] in in [sic] in [sic] the foreign service if you got sent to Guadalajara <,> it meant you weren’t doing very well <,> <ICE-GB:S1A-056 #54:1:A>

By concentrating on the individual subclasses of get constructions in ICE-GB, it has been shown that the preference for responsible over non-responsible subjects applies to pseudo, adjectival, idiomatic and reflexive get-constructions, while responsible subjects are unexpec-
tedly outnumbered by non-responsible ones in *get*-passives and semi *get*-constructions, as can be seen in Table 2.

<table>
<thead>
<tr>
<th>TYPE OF CONSTRUCTION</th>
<th>RESPONSIBLE</th>
<th>NON-RESPONSIBLE</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>get</em>-passives</td>
<td>7 (17.07%)</td>
<td>34 (82.93%)</td>
</tr>
<tr>
<td>semi <em>get</em>-constructions</td>
<td>5 (38.46%)</td>
<td>8 (61.54%)</td>
</tr>
<tr>
<td>pseudo <em>get</em>-constructions</td>
<td>28 (96.55%)</td>
<td>1 (3.45%)</td>
</tr>
<tr>
<td>adjectival <em>get</em>-constructions</td>
<td>31 (73.81%)</td>
<td>11 (26.19%)</td>
</tr>
<tr>
<td>idiomatic <em>get</em>-constructions</td>
<td>35 (97.22%)</td>
<td>1 (2.78%)</td>
</tr>
<tr>
<td>reflexive <em>get</em>-constructions</td>
<td>38 (97.44%)</td>
<td>1 (2.56%)</td>
</tr>
</tbody>
</table>

Table 2. Distribution of *get* + *Ven* according to subject responsibility.

As far as the animacy features of the subject are concerned, *get* constructions as a whole appear with an animate\(^3\) subject in almost 87% of the cases (e.g., (24) below) and this predominance is reflected in each of the six subclasses (see Table 3). There are, however, 28 examples (13%) where the subject is inanimate, which argues against Lakoff’s (1971: 154-155) oft-quoted claim that *get*-passives cannot take inanimate subjects. In most cases where the subject of the *get* construction is inanimate, some human entity associated with it retains responsibility, is involved in the action, or is affected by the results of the event (e.g., (25)), although there are cases in which the subject-referent does not stand in a direct relationship to any person, as in example (26), where the natural phenomenon of heat is not subject to human influence:

---

3 Practically all of the examples are human animates, with one single instance (the example below) where the referent of the subject is racing horses:

So they swing into the home straight and they’re coming down *<unclear-words>* stand side and *they’re getting sorted out* Indian file with Lord Chalmer setting the pace and a good one too to [sic] in second place *<,> Gwecko Solo <ICE-GB:S2A-006 #66:2:A>*
From prototypical to peripheral: ‘get + Ven’ in contemporary BrE

<table>
<thead>
<tr>
<th>TYPE OF CONSTRUCTION</th>
<th>ANIMATE</th>
<th>INANIMATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>get-passives</td>
<td>34 (68%)</td>
<td>16 (32%)</td>
</tr>
<tr>
<td>semi get-constructions</td>
<td>13 (100%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>pseudo get-constructions</td>
<td>28 (96.55%)</td>
<td>1 (3.45%)</td>
</tr>
<tr>
<td>adjectival get-constructions</td>
<td>36 (80%)</td>
<td>9 (20%)</td>
</tr>
<tr>
<td>idiomatic get-constructions</td>
<td>36 (100%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>reflexive get-constructions</td>
<td>40 (95.24%)</td>
<td>2 (4.76%)</td>
</tr>
</tbody>
</table>

Table 3. Distribution of get + Ven according to subject animacy.

(24) He [sic] got really pissed off when we were watching Back To The Future Two <ICE-GB:S1A-006 #86:1:A>
(25) But the moment you cross the lines <,> the shutters go up <,> your business gets nicked <,> <ICE-GB:S1A-027 #84:1:B>
(26) Note that an awful lot of heat gets gathered round the world and passed up in our uhm [sic] neck of the woods which keeps our climate mu [sic] much milder than it would otherwise be <,> <ICE-GB:S2A-043 #83:1:A>

4.2.5. Summary of findings

The overall findings of the analysis are summarised in Table 4:

<table>
<thead>
<tr>
<th>TYPE OF CONSTRUCTION</th>
<th>NO AGENT PHRASE</th>
<th>ANIMATE SUBJECT</th>
<th>ACTIVITY VERB</th>
<th>RESPONSIBLE SUBJECT</th>
<th>ADVERSATIVE SEMANTIC NUANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>get-passive</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>semi get-constructions</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>pseudo get-constructions</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>adjectival get-constructions</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>idiomatic get-constructions</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>reflexive get-constructions</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
</tbody>
</table>

Table 4. Overview of get + Ven characteristics.
5. Conclusions

The preceding analysis and discussion of get constructions in contemporary spoken British English allows us to confirm some of the syntactic, semantic and pragmatic characteristics attributed to central get-passives in the literature (cf. Quirk et al. 1985; Collins 1996; Carter/McCarthy 1999; Huddleston/Pullum et al. 2002). For example, get constructions occur primarily without an overtly expressed agent by-phrase. More interestingly, however, many of these features apply only to some get constructions and not to others, which would seem to confirm the idea that get constructions form a gradient with different degrees of prototypicality.

Regarding the type of lexical verb, not all get constructions take activity verbs, since semi and adjectival get-constructions tend to take stative past participles or adjectival -ed forms, and make considerable use of mental verbs. From a semantic perspective, only central get-passives and adjectival get-constructions tend to code unfortunate situations for the subject-referent, while favourable consequences for the subject are preferred in the other constructions (semi, idiomatic, and reflexive get-constructions). These situations, whether beneficial or adversative, are experienced most frequently by an animate and human subject throughout all the subcategories. Also, in most of the subclasses, the animate subject-referent carries some kind of responsibility for initiating the action described. However, the well-known feature of central get-passives, that they involve some responsibility on the part of the subject, only applies to non-central members of the category. In this respect, get-passives and semi get-constructions bear more similarities to prototypical be-passives in having as subject a non-responsible patient of the action. Further research into get passive constructions would help to confirm these tendencies.
References


Brown = Francis/Kučera (1964).


Hatcher, Anna G. 1949. To Get/Be Invited. Modern Languages Notes. 64/7, 433-446.


THOMAS EGAN

Encoding ‘throughness’ in English and French

Abstract

In this paper I discuss ways in which the concept which, for want of a better word, we may describe as ‘throughness’ is coded in English and French. The concept of ‘throughness’ is a relational one, the basic spatial sense of which involves a process whereby an object (a trajector in the terminology employed by cognitive linguists) moves along a path comprising an entrance to, traversal of, and exit from some sort of container. This study compares codings in English and French of this concept as it is instantiated in translations of the same Norwegian texts. Special attention is paid to tokens coding motion events. The data for the study are taken from the Oslo Multilingual Corpus.

1. Introduction

In his book Seeing through Multilingual Corpora, Stig Johansson touched on the status of tertium comparationis in contrastive studies. He wrote: “Much discussion in contrastive analysis has revolved around the question of the tertium comparationis, i.e., the background of sameness against which differences can be viewed and described” (Johansson 1997: 39). There is no doubt that any contrastive corpus linguist who takes translation equivalence as evidence of semantic equivalence is working on the overt or tacit assumption that there

1 Author’s email address: Thomas.Egan@hihm.no; affiliation: Hedmark U. College, Norway. The author would like to thank the editors and two anonymous reviewers for their valuable and constructive comments.
exists a viable tertium comparationis in the form of a meaning component common to both the source expression and its translation.

Krzeszowski (1990: 25) employs the term ‘2-text’ to refer to texts in either parallel or translated corpora. The availability of multilingual corpora, such as the Oslo Multilingual Corpus (OMC), allows us to operate with the concept of the ‘3-text’, with expressions in a source language serving as potential tertia comparationis for their translations into two other languages. In the present study I operationalise the notion of tertium comparationis in 3-texts in an investigation of how the notion of ‘throughness’ is encoded in English and French, comparing translation equivalents in English and French of the Norwegian preposition gjennom, which codes various sorts of ‘throughness’ relationship, both spatial and figurative/metaphorical. In §2 I introduce the corpus and give some reasons why I consider such a corpus to be suited for this sort of study. Section 3 describes in detail the notion of ‘throughness’, especially as it is coded by the Norwegian preposition gjennom, translations of which comprise the data for the study. Section 4 investigates English and French strategies for coding the main types of ‘throughness’ identified in §3. Finally, §5 contains a summary and conclusion.

2. Multilingual corpora as sources of tertia comparationis

The linguist wishing to compare or contrast lexemes or constructions in two or more languages must first establish some commonality on which to base his or her comparison. If two expressions have nothing whatsoever in common, one has no grounds for comparing them. For instance, it makes little sense to compare the meanings of two such nouns as apple and university. As Krzeszowski (1990: 15) puts it:

All comparisons involve the basic assumption that the objects to be compared share something in common, against which differences can be stated. This common platform of reference is called tertium comparationis. Moreover, any two or more objects can be compared with respect to various features and, as
a result, the compared objects may turn out to be similar in some respects but different in others.

Expressions in two languages may be similar lexically but differ semantically, as in the case of false cognates, such as English library and French librairie. In a case like this, the basis for comparison, the tertium comparationis, is the form and etymology of the two nouns. This comparison may be said to be semasiologically motivated. Another semasiological approach to the study of relational expressions in different languages involves the use of parallel or translation corpora. Schmied (1998), for instance compared the English preposition with with the German equivalent in a subcorpus of the Chemnitz English-German Translation Corpus. Cosme/Gilquin (2008) compared with with French avec, using material from the Poitiers-Louvain Échange de Corpus Informatisés. Both of these studies have taken lexical items as their starting point and investigated the extent of semantic equivalence between the forms. The present study, while it is also concerned with relational expressions, is onomasiologically motivated, comparing expressions coding ‘throughness’ in two languages. The tertium comparationis in this case is therefore semantic/pragmatic, rather than syntactic equivalence. The question that immediately springs to mind is how we are to go about identifying the semantic tertium comparationis that would serve to legitimise our comparison. One frequently employed method for comparing codings of the same meanings in different languages is to show speakers of various languages picture stories or video snippets and to get them to describe the actions they see in their own words. The most well-known examples of this approach are the many studies based on The Frog Story (see Berman/Slobin 1994). In the present study, on the other hand, the guarantor of semantic equivalence is the original item in a 3-text corpus, which is translated into the two items being compared.

The OMC is actually a 4-text corpus, in that it contains long extracts from 5 Norwegian novels translated into English, French and German. For the present study I only looked at the English and French

---

2 See <http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus>.
translators as my purpose was to compare these two languages. Apart from the fact that it provides the investigator with a *tertium comparationis*, the 3-text corpus has two other advantages over a 2-text corpus. In the first place, it allows for the comparison of identical text types, in that both the texts being compared are translations, whereas working with a 2-text translation corpus one is comparing an original text with a translation. In the second place, the 3-text corpus allows us to calculate the overlap between equivalent expressions in the two languages being compared. For example, Figure 1, from Egan (2013), shows the overlap between English *between* and French *entre* in tokens coding ‘betweenness’ as coded in Norwegian by phrases containing the preposition *mellom*.

![Figure 1](image.png)

**Figure 1.** The overlap between *between* and *entre* in coding ‘betweenness’ (Egan 2013).

Figure 1 shows that 79% of the tokens of ‘betweenness’ in the material were coded by *between* in English and 74% by *entre* in French. More importantly, it shows that 56% of tokens were coded by both these prepositions. It is impossible to arrive at this measure of overlap, which may be of use to both language teachers and compilers of automatic translation programmes, using a 2-text corpus.
3. The semantic field of ‘throughness’

Although the comparison of codings of ‘throughness’ in English and French in §4 is onomasiologically based, it takes as its tertium comparationis codings in Norwegian of the concept of ‘throughness’ by means of the preposition/particle gjennom. The main definition of gjennom in Norsk Ordbok, the closest Norwegian equivalent to the OED, is:

[U]sed about a movement or something perceived as motion which takes place in (within, surrounded by) that which is encoded by the landmark from one end or side all the way to the other, containing the whole landmark from start to finish; (in) from the one side or end and (out) to the other. (Vikør 2002; my translation)

This definition is very similar to standard definitions of the prototype of through, as described for example by Leech (1969: 181), Lindstromberg (1998: 31), Lee (2001: 49) Tyler/Evans (2003: 219) and Egan (2012: 44). There may well be relationships or processes which are conceived by Norwegian speakers as involving an element of ‘throughness’ that would not be so conceived by speakers of English and French. This has, however, no consequences for the present study, which is solely concerned with comparisons of English and French. If some of the tokens being compared do not correspond to the concept of ‘throughness’ as normally coded in one or other or indeed both of these languages, evidence for this mismatch will be provided by the English and French expressions themselves.

There are in all 322 tokens of gjennom in the OMC. The phrases containing eight of these are not translated into French and three are not translated into English. As two phrases are not translated into either language, this leaves us with 313 tokens to compare. These 313 tokens were assigned to eight main semantic categories, two of which are spatial, either Motion, as in example (1) below, or Location, as in (2). The eight categories are not taken from a dictionary or reference work, but were established by the author on the basis of the actual tokens in the material. The optimal degree of granularity for such a
classification will always be a matter of debate. One could, for instance, divide all tokens into three broad categories of spatial predications, temporal predications and ‘other’ predications. On the other hand, each token is in certain respects *sui generis*. In the present case I have chosen to group together tokens that seemed to be used to express similar meanings and that were sufficiently distinct from other senses as to warrant establishing a separate category. For the convenience of the reader who is not familiar with Norwegian, the English and French versions of each token will be cited rather than the Norwegian originals, which are listed in an appendix. The expressions which correspond to *gjennom* in the original texts are written in italics. The codes refer to the texts, the final two letters *TE* and *TF*, standing for ‘English translation’ and ‘French translation’, respectively.

(1a) Latour ran *through* the streets after the Marquis’ carriage. (NF1TE)
(1b) Latour courut derrière le carrosse *à travers* les rues de la ville. (NF1TF)
(2a) From that point of view, it would have been better for the date line to run *through* London or Paris. (JG3TE)
(2b) Cela aurait été différent si la ligne de changement de date était passée *par* Londres ou Paris. (JG3TF)

All six tokens of Location ‘throughness’ in my material resemble (2) in coding what Langacker (1991: 217) calls “subjective motion”.

Perception tokens, as in (3), resemble Motion tokens in so far as lightwaves or soundwaves move from the object of perception along a path to the sense organ. However, we normally conceive of sense impressions as immediate, not as something moving along a path. Moreover, if we do think about them in terms of a path, we are likely to think of progress along it as being in the opposite direction to the scientifically correct one, i.e., from the perceiver to the perceived.

(3a) Mother Karen looked at Dina *through* her monocle and a veil of goodwill. (HW2TE)
(3b) Mère Karen considéra Dina *à travers* son monocle et un voile d’infinie bienveillance. (HW2TF)

Just as ‘throughness’ may be conceived of in terms of a path through space, it may also be conceived of in terms of a path through time. Example (4) exemplifies this temporal sense.
There are a handful of tokens where expressions of ‘throughness’ refer to a means or instrument, by means of which a trajector is helped in negotiating a path towards a landmark. Example (5) exemplifies this sense.

(5a) Unfortunately I feel unfree towards her and become virtually dumb as soon as we do not communicate with each other through music. (BHH1TE)
(5b) Malheureusement, dès que nous cessons de communiquer par le biais de la musique, je me sens tout gêné en face d’elle, au point d’en rester muet. (BHH1TF)

Closely related semantically to the Means sense is a usage whereby the landmark of the predication of ‘throughness’ represents a sort of conduit, a channel or medium, as in (6).

(6a) She sent get-well wishes through him. (HW2TE)
(6b) Elle le chargeait de transmettre des voeux de meilleure santé. (HW2TF)

A number of tokens of gjennom occur in phrases that are idiomatic in Norwegian. These may or may not correspond to idiomatic phrases in other languages. They have been categorised separately. Example (7), which contains translations into English and French of the Norwegian multi-word verb gå gjennom, meaning ‘examine’ (literally ‘go through’), may serve as a typical example of these tokens.

(7a) He went over it in detail, deciding on the instruments he would use and the incisions he would make. (NF1TE)
(7b) Il en récapitula toutes les opérations, prévit les points d’incision, les instruments indispensables. (NF1TF)

These seven categories account for just 85% of all tokens of gjennom. The remaining 15% are metaphorical extensions of the Motion sense, as in (8), in which no physical motion is involved. I decided to label this category Other (metaphorical), rather than simply Metaphorical,
since the latter term could equally be applied to other non-spatial senses, such as the temporal sense.

(8a) He ploughed through page after page of longwinded prose. (NF1TE)
(8b) Il avança, page après page, à travers les interminables passages. (NF1TF)

Table 1 contains details of the incidence of all eight senses among the tokens translated into both languages in the corpus, in descending order of frequency. The table shows that almost half of the tokens instantiate the Motion sense. How Motion ‘throughness’ is coded in English and French will be the main focus of attention in the next section.

<table>
<thead>
<tr>
<th>SEMANTIC FIELD</th>
<th>NUMBER OF TOKENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>146</td>
</tr>
<tr>
<td>Perception</td>
<td>62</td>
</tr>
<tr>
<td>Other (metaphorical)</td>
<td>46</td>
</tr>
<tr>
<td>Time</td>
<td>24</td>
</tr>
<tr>
<td>Idiom</td>
<td>13</td>
</tr>
<tr>
<td>Medium</td>
<td>8</td>
</tr>
<tr>
<td>Means</td>
<td>8</td>
</tr>
<tr>
<td>Location</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
</tr>
</tbody>
</table>

Table 1. Types of ‘throughness’ encoded by gjennom.

4. English and French strategies for coding ‘throughness’

In studies based on translation corpora it is common practice to distinguish between congruent and divergent translations. As I am not concerned in this study with translations per se, but rather with correspondences between two sets of translated tokens, I employ instead the terms prepositional and non-prepositional. In the case of prepositional codings, I draw a further distinction between those translations that employ the most frequently used prepositions, through in English
and à travers in French, and other prepositions. Total figures for these three forms of translation into both languages are given in Figure 2.

![Figure 2. Forms used in English and French to translate expressions containing Norwegian gjennom.](image)

We can see in Figure 2 that, whereas prepositional forms are much more common in English than non-prepositional ones (274 tokens compared to 39 tokens), and through much more common than all other prepositions put together (228 compared to 46), in French prepositional and non-prepositional constructions are equally common, represented by 158 and 155 tokens, respectively. Moreover, just as through stands out from the other English prepositions in terms of frequency of occurrence, the verb traverser stands out among the French non-prepositional forms. Figure 3 shows the correspondence in codings between the two languages, with traverser distinguished from the other non-prepositional forms.
In Figure 3 the four French forms are distinguished on the horizontal axis, with the three English ones being indicated by various patterns. In the first column we can see how the individual tokens coded in French by à travers correspond to English translations by through, by other English prepositions and by English non-prepositional constructions. Thus, just over 60 tokens correspond to through and just a handful to other prepositional and non-prepositional constructions. Figure 4 shows the overlap between through and à travers in coding ‘throughness’. It may be compared to Figure 1 which shows the overlap between between and entre in codings of ‘betweenness’.
Figure 4 is not drawn to scale. If it were, the size of the \( \text{à travers} \) ellipse would be much smaller. Nevertheless, it contains some useful information on the extent to which the two prepositions may be substituted for one another. This relationship is obviously not mutual. While \( \text{through} \) corresponds to \( \text{à travers} \) in 87.5% of all cases in which the latter is used in French, the opposite is only the case in 29% of cases. In the following four sections I will look at correspondences on the level of the four most common types of ‘throughness’, starting with the Motion sense.

4.1. Coding Motion ‘throughness’

Figure 5 contains the raw figures for the overlap between three English and four French means of coding Motion ‘throughness’. It shows that all four forms of encoding in French correspond to translations by \( \text{through} \) in English. The proportion of the four sets of columns resembles those depicted for the ‘throughness’ relation in general in Figure 3, with the exception of a somewhat higher incidence of tokens coded by the verb \( \text{traverser} \).
To begin with tokens coded by the two propositions *through* and *à travers*, there are in all 23 tokens coded by both. In 17 of these the manner of motion is coded by the verb in French, as in (3b) and (9b). These French tokens resemble their English counterparts insofar as over 75% of the latter code manner of motion in the verb phrase. The remaining six French tokens resemble (10b) in containing a general verb of motion.

(9a) We began to walk slowly *through* the galleries, and up to the first floor. (JG3TE)
(9b) Nous *déambulâmes* un moment *à travers* les salles, puis montâmes au premier étage. (JG3TF)
(10a) I suggested I could walk with him *through* the Retiro Park. (JG3TE)
(10b) Je proposai de l’accompagner *à travers* le Retiro. (JG3TF)

Figure 5 shows that French translators actually prefer to employ other prepositions rather than *à travers* to code Motion ‘throughness’. I mentioned in §3 that the basic sense of ‘throughness’, as coded by Norwegian *gjennom*, is a path relation involving subcomponents of ‘entrance to’, ‘crossing of’ and ‘exit from’ an area or container. Only the central portion of the path, denoting the crossing of the area or container, is salient in all tokens in Norwegian. (Note that in the translation of the definition of *gjennom* in §3, the prepositions *in* and *out* are enclosed in brackets.) We should therefore expect to find tokens in
both English and French that make no reference to ‘entrance into’ or ‘exit from’ this area or container. Moreover, since prepositions are not generally used in French to code boundary-crossing activities (see, e.g., Cappelle 2012: 177), it comes as no surprise that French translators often focus on the central section of this path or the space within which the motion takes place. There are thus as many as 30 examples of the preposition *dans* in the French texts, as in (12b) and (13b), as opposed to just two of *in* in the English ones. And in the case of one of these, cited as (11a), the *in* phrase could be interpreted as a post-modifier of the noun ‘path’, with the actual path ‘through’ coded by *along*. The reason I have not analysed it as such is that the landmark of the preposition *gjennom* in the Norwegian original text is the woods, not the path.

(11a) I remember I was running along a path in the woods [...] (JG3TE)
(11b) Je me revois encore traverser en courant la forêt [...] (JG3TF)
(12a) When he walked through town [...] (BHH1TE)
(12b) Lorsqu’il déambulait dans la ville [...] (BHH1TF)
(13a) During the summer, Dina began wandering about the house. (HW2TE)
(13b) Dina recommença à circuler dans la maison cet été-là. (HW2TF)

In (12b), as in (9b), the manner of motion is coded by the verb. However, while in (9b) the path of motion is coded in a prepositional phrase, in (12b) it is the location within which the path is situated that is coded in this way. This combination of a verb coding manner and an adverbial coding location is found in 23 of the 30 tokens containing *dans*.

If an area to be crossed is very small, or even two-dimensional, the option of concentrating on the central portion of the path is not available to the language user. In such cases, containing what Aurnague (2000) refers to as an intrinsically medial space, French speakers may choose to code the path in the verb phrase and encode the point of boundary crossing by the preposition *par*, as in (14b)-(16b).

(14a) Then crawled through the open window. (HW2TE)
(14b) Elle sortait ensuite par la fenêtre. (HW2TF)
Since then, the room can be entered only through the courtyard [...] (BHH1TE)

Depuis, il fallait passer par la cour pour accéder à la pièce [...] (BHH1TF)

One spring a duck entered the cookhouse through the open door [...] (HW2TE)

Une année, une mère eider entra par la porte ouverte du fournil [...] (HW2TF)

There are 13 instances of par in the French texts, but not a single one of English by, although the latter would certainly be deemed acceptable in contexts like (15a) and (16a), in which the path is coded by the verb. It would not, however, be possible to substitute by for through in (14a), in which the verb codes manner of motion, rather than path. That is, substituting by for through would result in an idiomatic sentence, but the direction of the path would be orthogonal to the original one.

Figure 5 shows that there are far more instances of non-prepositional codings of Motion ‘throughness’ in French than in English. We have already seen in (14b)-(16b) some examples in which the path is coded by a combination of verb and preposition, with the direction of the boundary crossing being coded in the verb and the boundary itself in the preposition phrase. There are also 26 tokens in French in which the path is coded by the verb alone, as in (17b) and (18b). There are only 6 such examples in English, one of which is cited as (17a).

He climbed a staircase, traversed corridors that smelt of nutmeg [...] (NF1TE)

Il grimpa un escalier, traversa les couloirs où flottait un parfum de musc [...] (NF1TF)

A steady stream of pupils, moving on foot and by bike, cut through the gates in the fence [...] (KF1TE)

À pied ou en vélo, un flot interrompu d’élèves franchissait les portillons ménagés dans la clôture [...] (KF1TF)

Table 2 contains details of the codings of Motion ‘throughness’ in English and French, with respect to the extent to which path and manner in motion are coded in verb phrases, in prepositional phrases, or a combination of the two. Verb phrases that code motion but are neutral with respect to both manner and path, such as non-deictic come and
venir are not included in the table. Also excluded are cases where the manner or path is coded by non-prepositional adverbials.

<table>
<thead>
<tr>
<th></th>
<th>ENGLISH</th>
<th></th>
<th>FRENCH</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VP</td>
<td>PP</td>
<td>VP + PP</td>
<td>VP</td>
</tr>
<tr>
<td>MANNER</td>
<td>111</td>
<td>0</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>PATH</td>
<td>6</td>
<td>128</td>
<td>1</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 2. Coding of Manner and Path in Motion tokens in English and French.

The figures in Table 2 show that the evidence of the texts in the present study conforms to the generally accepted view that English is indeed a satellite-framed language. The picture for French is more mixed with respect to Talmy’s (2000: 221) typological distinction, as has been pointed out by Kopecka (2006) and Pourcel/Kopecka (2005). While obviously verb-framed to a much greater extent than English, the tokens in this study nevertheless display many more instances of satellite-framing than one might have expected if French were a pure verb-framed language.

4.2. Coding Perception ‘throughness’

Figure 6. Correspondences between English and French codings of Perception ‘throughness’.
Figure 6 contains the raw figures for the overlap between three English and four French means of coding Perception ‘throughness’. The evidence presented in the figure shows that codings of Perception ‘throughness’ differ markedly from those of Motion ‘throughness’ presented in Figure 5, particularly with respect to the number of tokens coded by both through and à travers, as in (19)-(21).

(19a) There was my master’s hut visible through the canopy of leaves. (NF1TE)
(19b) J’entrevois la baraque de Léopold à travers le feuillage. (NF1TF)
(20a) Sea gulls shrieked to them through the open windows. (HW2TE)
(20b) Les mouettes criaient à travers les fenêtres ouvertes. (HW2TF)
(21a) He sensed Dina’s aroma even through the odors of herring barrels [...] (HW2TE)
(21b) L’odeur de Dina lui parvenait à travers les tonneaux de harengs [...] (HW2TF)

There is an obvious reason why French should be more likely to code the path of perception in a preposition phrase rather than the verb phrase. In predications of perception the verb is generally employed to code the mode of perception, i.e., the particular sense involved, as in (19b). Indeed, it may code both mode and manner, as in (20b). However, even in cases where the verb is neutral with respect to the mode of perception, as is the case with parvenir in (21b), we still find instances where the path is encoded by a preposition phrase.

À travers is used to code 57% of the Perception tokens in French. The remainder are rendered by both other prepositional and non-prepositional constructions, exemplified here by (22b) and (23b) respectively.

(22a) Through the red material of her dress I glimpsed the black bikini-top. (JG3TE)
(22b) Sous l’étoffe légère de sa robe, je pouvais deviner le haut noir de son bikini. (JG3TF)
(23a) The weather was hot, it had cleared up after some heavy morning showers, but now I felt a cold shiver pass through my body. (JG3TE)
(23b) Il faisait chaud. Après quelques violentes averse, le ciel s’était enfin éclairci. Pourtant, je sentis un frisson me parcourir le corps. (JG3TF)

In (22b) the French writer first locates the bikini top beneath the woman’s dress and then states that the subject could see it, thus
implying that the dress material was transparent. The English writer, on the other hand, opts for coding this transparency more directly. In (23) both writers opt for forms that we also find in motion tokens, the feeling perceived in this case being one that traverses the organ of perception, rather than being merely registered by it.

4.3. Coding Time ‘throughness’

Figure 7 contains the raw figures for the overlap between three English and four French means of coding temporal ‘throughness’. Time tokens differ markedly from both Motion and Perception tokens in that they exhibit no overlap whatsoever between codings by through and à travers. Indeed there is only one token of the latter in the French texts, and only one token of traverser. The English and French texts do, however, resemble one another in so far as both contain a large number of tokens in which Time ‘throughness’ is coded by prepositions other than through and à travers. These are exemplified here by (24)-(26).

![Figure 7. Correspondences between English and French codings of Time ‘throughness’](image_url)
(24a) I believe that crafty rogue can smuggle a bit of anything up and down the centuries. (JG3TE)
(24b) Je crois que ce petit malin peut faire passer en douce d’un siècle à l’autre, en avant ou en arrière, tout ce qui lui chante. (JG3TF)
(25a) Niels had operated very cleverly for many years. (HW2TE)
(25b) Niels avait montré une grande intelligence pendant plusieurs années. (HW2TF)
(26a) He probably overtaxed himself in his youth, for in his mature years his vigour declined so quickly [...](BHH1TE)
(26b) Sans doute avait-il trop forcé au cours de ses jeunes années car, durant sa maturité, sa santé déclina si vite [...] (BHH1TF)

Almost half of the French tokens containing prepositions have durant, as in (26b). No single preposition stands out to this extent in English. In addition to those employed in (24)-(26), we find tokens containing during, over and against.

Both languages also contain a repertoire of non-prepositional constructions that may be drawn upon to profile the prolonged existence of an activity or state throughout a period of time. Some of these are exemplified in (27)-(29).

(27a) But the photo he’d taken had already hounded Ana all her life. (JG3TE)
(27b) Mais ce portrait avait, en réalité, poursuivi Ana toute sa vie. (JG3TF)
(28a) (1) He had been a member of the National Union since 1934 and paid his membership fee every year. (BHH1TE)
(28b) (1) Il avait été membre du N.S. de 1934 à 1945 et s’était régulièrement acquitté de sa cotisation. (BHH1TF)
(29a) I found the “Nine O’clock Hour” to be a great help during the difficult period of transition from sleep to waking [...] (KF1TE)
(29b) Mes préférences allaient au concert de midi et à l’émission fourre-tout de neuf heures du matin qui m’aidait à franchir la difficile période transitoire [...] (KF1TF)

Both the English and French writers in (27) adopt a similar form to code an expression in Norwegian that may be more literally translated into English as “throughout her life”. In (28) they adopt slightly different strategies, the English writer indicating that the fees were paid throughout the subject’s period of membership by stating that they were paid each year, while the French writer implies the same constancy of payment by stating that they were paid on a regular basis. In
(29a) the English writer opts for a prepositional construction similar to that chosen by the French writers in (25b) and (26b), while the French writer chooses to code the ‘throughness’ in the verb *franchir*, which we also find used to code Motion tokens.

### 4.4. Coding Other (metaphorical) ‘throughness’

Figure 8 contains the raw figures for the overlap between three English and four French means of coding Other (metaphorical) ‘throughness’. One striking difference between the data in Figure 8 and those in Figure 7 is the number of tokens coded by *à travers* and *traverser*. Moreover, every single one of these tokens in Figure 8 corresponds to one containing *through* in the English texts. Examples (30)-(32) exemplify this overlap.

![Figure 8. Correspondences between English and French codings of Other (metaphorical) ‘throughness’.

(30a) [...] to guide her with a gentle hand *through* life’s ups and downs. (BHH1TE)
(30b) [...] de la mener précautionneusement *à travers* toutes les vicissitudes de la vie. (BHH1TF)
(31a) They flow *through* my head like water. (HW2TE)
(31b) Ils *traversent* ma tête comme de l’eau. (HW2TF)
(32a) He never said much at all, just watching the anatomist with eyes that seemed to burn right through his pale-coloured gown. (NF1TE)
(32b) Il parlait peu, se contentant de suivre l’anatomiste avec un regard si enflammé qu’il semblait traverser sa redingote claire. (NF1TF)

Example (30) differs from (31) and (32) in that the former codes (metaphorical) Caused Motion and the latter (metaphorical) Self-Motion. In cases of Caused Motion, the mode of causation is normally coded by the verb in both French and English, leaving it in both languages up to an adverbial to code the path. In cases of Self-Motion, there is no need for the verb to code a causative element, leaving it free to code the path, as in (31b) and (32b). Very occasionally French verbs may code manner as well as path, as in (33b). More often, French writers rephrase the entire predication of ‘throughness’, as in (34b) and (35b), and as the English writer does in (35a).

(33a) Envious glances darted through the underbrush [...] (BHH1TE)
(33b) Et d’envieux regards perçaient la forêt de petits arbres [...] (BHH1TF)
(34a) Sometimes it crossed his mind that this was unnatural, animal-like. (HW2TE)
(34b) De temps en temps, il avait le sentiment que cela était bestial et contre nature. (HW2TF)
(35a) [...] even if her mind was in turmoil. (NF1TE)
(35b) [...] même quand elle était assaillie par des flots de pensées. (NF1TF)

The Norwegian original of (34) may be translated literally ‘went through his brain’. The English translator codes the throughness in the verb ‘cross’. The French translator also captures the sense of the predication, but he profiles it differently by coding the ‘thinker’ as subject and possessor of the thought in question. In (35) both translators opt for a reprofiling of the original predication which, literally translated into English, goes something like ‘thoughts rushed through her’.
5. Summary and conclusions

In this paper we have seen how predications of ‘throughness’, as coded by the preposition gjennom in Norwegian, are rendered in English and French. The data for the study were taken from the OMC, which contains 313 tokens of gjennom translated into both English and French. Compared to the way in which both of these languages code ‘betweenness’ (see Egan 2013), there is much less similarity between English and French in their encoding of ‘throughness’. I distinguished between eight semantic types of predication. There are considerable differences in the ways the two languages code these various types of ‘throughness’. Indeed, the differences between the two languages in their encodings of these various subsenses are all statistically significant. Moreover, there is a only an overlap of 21% between English and French in their use of what might be considered the default prepositions for coding ‘throughness’, through and à travers. French employs prepositional constructions much less than English to encode ‘throughness’, only 56% of the time, in fact. This ratio is similar to that adduced by Paulussen (1999) in a 3-text study of on/up in English and sur in French.

Of the eight main types of predication, four are represented by more than 20 tokens among the 313 tokens in the study. The relevant domains of predication are Motion, Perception, Time and Other (metaphorical). All four were discussed in §4. Perception tokens resemble one another in English and French more than the other three main types, with 48% encoded by both through and à travers. In such predications the actual physical organ involved in the act of perception is often coded by the verb, leaving it up to an adverbial to code the path, if so desired. Time tokens, on the other hand, resemble one another in English and French in the number of tokens containing prepositions other than through and à travers. In fact, not a single token is encoded by both. Motion tokens, which account for almost half the total in the corpus, differ greatly with respect to encoding by through (75%) and à travers (18%). The latter is actually exceeded in number by dans (21%). Traverser accounts for 21% of the tokens and other
non-prepositional constructions 28%. The difference between the two languages with respect to Motion tokens is no doubt related to typological differences. However, it was pointed out in §4.1 that, whereas English behaves in my material as one would expect of a satellite-framed language (see Talmy 2000), the picture for French is much more mixed. While the majority of French tokens code path in the verb phrase, there is a sizable minority that code it in a prepositional phrase. Other (metaphorical) tokens are similar to Motion tokens as far as English is concerned (76% through). As for French, the main difference resides in the number of non-prepositional constructions other than those containing traverser. These account for 54% of all tokens, as opposed to just 15% of the English ones.

To sum up, I hope that the evidence presented in this paper about the various means of coding ‘throughness’ in English and French, and in particular on the degree of overlap in coding strategies employed by the two languages in the case of the various subsenses, may contribute to extending our understanding of the ways in which this particular relational concept is coded in these two typologically dissimilar languages.

References


**Appendix**

This appendix contains the Norwegian originals of all the cited examples.

(1) Latour løp etter vognen hans gjennom gatene. (NF1)
(2) Da hadde det vært bedre om datogrensen hadde gått gjennom London eller Paris. (JG3)
(3) Mor Karen så på Dina gjennom monokkel og et slør av velvilje. (HW2)
(4) Opp gjennom tidene gjorde pavestolen hva den kunne for å komme legenden til livs. (BHH1)
(5) Dessverre er jeg ufri overfor henne og blir nærmest stum så snart vi ikke meddeler oss til hverandre gjennom musikken. (BHH1)
(6) Hun overbragte god-bedring-ord gjennom ham. (HW2)
(7) Han gikk gjennom den i detalj, planla snittene, hvilke instrumenter han skulle bruke. (NF1)
(8) Han bladde og bladde gjennom de langdryge passasjene. (NF1)
(9) Vi begynte å gå langsomt gjennom galleriene, og opp i andre etasje. (JG3)
(10) Jeg foreslo at jeg kunne følge ham gjennom Retiro-parken. (JG3)
(11) Jeg husker jeg løp på en sti gjennom skogen [...] (JG3)
(12) Når han kom gående gjennom byen [...] (BHH1)
(13) Dina begynte å gå gjennom stuene denne sommeren. (HW2)
(14) Og så klarte ut gjennom det åpne vinduet. (HW2)
(15) Etter det har hybelen bare adkomst gjennom garden [...] (BHH1)
(16) Et år kom ei ærfuglmor seg inn i eldhuset gjennom den åpne døra og ordnet seg til i den store bakervinen. (HW2)
(17) Han skrittet opp en trapp, gikk gjennom muskatluktende ganger [...] (NF1)
(18) En stadig strøm av elever til fots og på sykkel, sneiet inn gjennom porter i gjerdet [...] (KF1)
(19) Gjennom løvverket skimtet jeg mesterens koie. (NF1)
(20) Måkene skrek inn til dem gjennom åpne vinduer. (HW2)
(21) Han hadde teven av Dina tvers gjennom sildetønner [...] (HW2)
(22) Gjennom det røde kjolestoffet kunne jeg skimte den sorte bikini-overdelen. (JG3)
(23) Varmt var det, for det hadde omsider klarnet opp etter noen heftige morgenskurer, men nå kjente jeg at det for en kuldegysning gjennom kroppen. (JG3)
(24) Jeg tror at den luringen kan smugle både det ene og det andre opp og ned gjennom århundrene. (JG3)
(25) Niels hadde brukt stor klokt gjennom flere år. (HW2)
(26) Trolig har han tatt for tunge løft i de unge år, for ut gjennom manndomsalderen gikk det så raskt tilbake med førligheten [...] (BHH1)
(27) Men bildet dvergen knipset hadde allerede forfulgt Ana gjennom hele livet. (JG3)
(28) 1. Han hadde vært medlem av Nasjonal Samling siden 1934 og betalt kontingent gjennom alle år. (BHH1)
(29) Jeg syntes gudhjelpemeg Ni-timen var en støtte gjennom den vanskelige overgangsperioden fra søvn til våken tilstand. (KF1)
(30) [...] å hjelpe og støtte henne og føre henne varsomt gjennom alle livets ujevnheter. (BHH1)
(31) De renner gjennom hodet mitt som vann. (HW2)
(32) Han sa ikke stort, bare fulgte anatomen med et blikk som formelig brente seg gjennom den lyse kappen hans. (NF1)
(33) Misunnelige blikk skjøt gjennom småskogen [...] (BHH1)
(34) Noen ganger før det gjennom hjernen hans at dette var dyrisk og unaturlig. (HW2)
(35) [...] selv om et ras av tanker for gjennom henne. (NF1)
If you would like to lead: on the grammatical status of directive isolated *if*-clauses in spoken British English¹

Abstract

The aim of this paper is to analyse so-called isolated *if*-clauses (e.g., *Okay if you’d like to get dressed now*; *If I’d somewhere to go, some friend’s room*) in contemporary spoken British English on the basis of data extracted from the *Diachronic Corpus of Present-day Spoken English* (Aarts/Wallis 2006). Isolated *if*-clauses have the same form as ordinary conditional clauses, but lack a superordinate matrix (hence the label ‘isolated’) and have the illocutionary force of polite requests, directives and wishes. The focus here is on the directive type. Aspects considered include modality and grammatical status, that is, whether such clauses should be seen as incomplete utterances marked by ellipsis or as actually constituting a minor clause type. Also discussed is the relation of such clauses to the phenomenon of insubordination (Evans 2007).

¹  Author’s email address: beatriz.mato@usc.es; affiliation: U. of Santiago de Compostela. For generous financial support I am grateful to the European Regional Development Fund and the following institutions: the Spanish Ministry of Economy and Competitiveness (grant FFI2011-26693-C02-01), the Autonomous Government of Galicia (grant CN2012/012), Spanish Ministry of Education (grant FPU AP2010-4570). I am also grateful to two anonymous reviewers and to Teresa Fanego and María José López-Couso for their feedback on an earlier version of this paper.
1. Introduction

The present paper intends to analyse the so-called directive isolated *if*-clauses in contemporary spoken British English, on the basis of data extracted from the *Diachronic Corpus of Present-day Spoken English* (*DCPSE*) (Aarts/Wallis 2006). I will first comment briefly on the existing literature on the topic (cf. §2). In §3 I will consider the modality and grammatical status of directive isolated *if*-clauses, that is, whether they should be seen as incomplete utterances marked by ellipsis or as actually constituting a minor clause type. I will provide corpus evidence that they are independent clauses, in view of the nature of the alleged missing material and its recoverability, and of the features they share with prototypical independent clauses. In turn, §4 discusses the stage of insubordination that directive isolated *if*-clauses seem to have reached. Finally, §5 summarises the conclusions and proposes some future lines of research.

2. Isolated *if*-clauses

The label isolated *if*-clause is used to refer to structures like those in (1) and (2) taken from Stirling (1999), in which an apparent conditional clause is issued on its own, without a corresponding apodosis or main clause:

(1) Okay *if you’d like to get dressed now.*
(2) If I’d *somewhere to go, some friend’s room.*

As far as English is concerned, Stirling (1999) points out that the *if*-clause in (1) expresses a polite request, that is, a directive, whereas (2) shows desiderative value, expressing a wish on the part of the speaker. These two uses of isolated *if*-clauses in English have been mentioned occasionally in the literature. It is worth mentioning, however, that isolated *if*-clauses can also express either surprise or a warning or
If you would like to lead: *directive isolated if*-clauses in spoken BrE

threat, as in the made-up examples (3) and (4). These uses make us consider this type of structures as a more functional and diverse resource than what can be deduced from the literature.

(3) *If it isn’t you again!*
(4) *If you (dare) break the window!*

In what follows, I will briefly report the presence of directive isolated *if*-clauses in the literature, paying special attention to the study of this type of clauses in Australian English carried out by Stirling (1999).

2.1. Isolated *if*-clauses in the literature

Isolated *if*-clauses have been commented on occasionally by several authors, e.g., Huddleston/Pullum et al. (2002: 945), who refer to conditional fragments as a type of conditional adjuncts that “can be used on their own, with the apodosis left unexpressed”, as exemplified below.

(5) *If only you’d told me earlier!*
(6) *Well, if it isn’t my old friend Malcolm Duce!*
(7) *If you’d like to move your head a little.*

The construction in (5) with the combination “*if only* + modal preterite indicating counterfactuality” expresses a regret with the implied meaning “how unfortunate you didn’t tell me earlier (because if you had done, things would have been better)”, the same use reported for our earlier example (2). In (6), in turn, the construction involves a fixed frame of the form *if it/that isn’t X*, which is used to express surprise at seeing X, so the example here itself conveys “it is my old friend Malcolm Duce”. The example under (7) is a further type of indirect directive: “please, move your head a little” (as said by doctor to patient); the missing apodosis is understood along the lines of “that would be helpful”.

Ford (1997) also identifies the possibility for conditional clauses to propose some action on the part of the hearer. As Ford (1997: 401) explains, directives and offers are places for special handling in
conversation since they involve an imposition on another person’s independence and freedom of choice. She also acknowledges that conditional clauses perform offers and other moves to make it clear that their realization depends on the addressee’s will, as in (8).

(8) You can come an’ sit’n talk with us if you want.

Ford states (1997: 405) that making an offer through a conditional clause is a way of dealing with the negative face wants of the addressee, the conditional allowing the hearer’s freedom of choice. As Traugott et al. (1986: 7) recognise, “it is invariably found that some sentences with the formal markers of conditionality are semantically and pragmatically only marginally conditional or not conditional at all”. This would be the case of both prototypical conditional clauses with directive meaning – the type of sentences from which isolated if-clauses seem to have emerged (cf. §4) – and isolated if-clauses, which do not convey conditional meaning, but rather serve other purposes, basically forming polite requests, offers and orders. In this respect, it has been shown that conditional clauses serve to the softening of the strong assertion that an imperative would imply. As Brown/Levinson (1987: 65-74) explain, orders, requests (i.e., the speaker’s expression of his/her desire that the hearer do or not an action) and offers (i.e., the speaker expecting that the hearer commits himself/herself to whether or not (s)he wants the speaker to do something for him/her) are acts that threaten the addressee’s negative face by expressing that the speaker intrudes in the hearer’s freedom of action. In this context, participants in a conversation will try to avoid these face-threatening acts or, at least, will try to employ some strategy to reduce the threat. Conditionals serve to the strategy of acts done on record (i.e., where the communicative intention behind the action is clear to participants) with redressive action, that is, an action that attempts to go against the potential damage the face of the addressee may suffer by modifying or adding some element that clearly expresses that it is not the speaker’s intention to do so. These acts having redressive action can be done both with positive or negative politeness, the latter consisting of acts oriented to maintain the hearer’s want of “territory and self-determination” (Brown/Levinson 1987: 70); in this type of acts, the
speaker avoids interfering with the hearer’s freedom of action. Face-threatening acts are redressed with the indirectness conditionals offer, since they provide the hearer with the possibility of choosing whether he/she wants or does not want to commit himself/herself to the act proposed by the speaker.

Ford/Thompson (1986) also comment on the uses of isolated if-clauses in order to express polite requests. They consider that it may be due to “a combination of the softening effect of hypotheticality and the fact that conditionals seem to imply an option with alternatives” (1986: 365). In many of the examples analysed by Ford/Thompson, a consequent clause is very difficult to isolate since this use of the conditional form is one of the least compatible ones with logical interpretation. The response from the interlocutor addressed often reflects the understood intent of the utterance, that is to say, the second speaker responds with assent:

(9) M: If you could get your table up with your new sketches just soon as this is over I would like to see you.
T: All right. Fine.

(10) M: But if you’ll call Irey over and get together with him on Tuesday or Wednesday, whenever you fellows are ready I’m ready.
J: Yes, all right, that’s fine.

Ford (1993: 49-50) also recognises the existence of if-clauses presenting options that occur unconnected to distinct main clauses, expressing an offer that is contingent on the hearer’s choice, as in (11) and (12) below.

(11) A: Tch! But if you wanna-uhm come in, an’ see.
B: Tch! I wouldn’t know where to look for her (hh) hhnh-hnhh!

(12) A: Well if you want me (to) give you a ring tomorrow morning.
B: Tch! Well y-you know, let’s, eh- I don’t know. I’ll see maybe I won’t be in.

B’s answers here reflect the interpretation of if-clauses as offers; A’s if-clauses are used in contexts where B seems to be hesitant about accepting the offer. The optionality that if-clauses convey seems to be used in sequences which involve hesitancy on the part of one participant. Speakers treat these autonomous adverbial clauses as complete
actions for themselves. Conditional clauses used without apodosis are used in these contexts since they encode options; in Ford’s words, “when one makes an offer [...] an if-clause is a workable format for suggesting the plan of action and at the same time displaying a recognition, or conceding to the fact, that the plan is contingent and the other party may prefer another option” (1993: 139). Speakers consider the if-clause alone as sufficient for the encoding of a polite offer and, in such contexts, it is not regarded as ungrammatical or incomplete, which Ford sees as an indicator of the progressive conventionalization of this type of clauses.

2.2. Stirling’s analysis of isolated if-clauses in Australian English

The most detailed discussion of directive isolated if-clauses in English was carried out by Stirling (1999) for Australian English. She examined the uses and features of these clauses extracting data from the General Practice Corpus, which consists exclusively of conversational dialogue recorded in Queensland in 1980 with the cooperation of The Royal Australian College of General Practitioners Family Medicine Program, and the Macquarie Dictionary Corpus. It consists of 20 million words covering a range of genres, though most of the data come from the written language (for further details see Stirling 1999: 274-275). The subset of genres chosen for Stirling’s study aimed to represent as conversational-like language as possible, taking into consideration the limitations that dialogue in plays and dialogue in third person thought in novels offer as an indication of actual usage.

All the examples Stirling reports from the General Practice Corpus were uttered by doctors in two phases of the consultation: in the concluding one when detailing the treatment and, less commonly, in the examination phase. Examples from the latter express the doctor’s request for the patient to move parts of his/her body, while examples from the treatment phase imply specification on further treatment the patient will have to perform without the supervision of the doctor. These examples usually relate to acts previously mentioned in the consultation, functioning thus as a reinforcement. Both situations require the patient’s cooperation to fulfil the acts being requested by
the doctor: in the first case, a movement of a part of the patient’s body and, in the second, the fulfilment of the recommendations made by the doctor, in which the use of isolated if-clauses seems to give the patient the chance to decide whether or not to fulfil the act.

In her discussion of directive isolated if-clauses, Stirling (1999: 278) states that “the speaker uses the isolated if-clause with the illocutionary force of a directive, most frequently a request (but in some cases more appropriately categorised as a suggestion)”. Consider example (13) below:

(13)  D: Deep breaths…If you’d like to move your head a little. Thank you.

In this type of isolated if-clauses, the subject is normally a second person one and, in those cases where a first person is found, the utterance has the force of a request for permission, as in (14) below:

(14)  MR. PETTIT: If I could answer that, Mr. Commissioner. My job in 1988 – and it took most of that year – was to consult with communities all around Victoria regarding the concept of the community justice panels […].

Stirling (1999: 279) points out that “directives naturally refer to a future action, and so it is not surprising that the simple past tense only rarely occurs […] instead one finds either the present form of the main verb, the present tense form of the modal can, or the past form of a modal”, as in (13) and (14) above. It must be noted, however, that past tense forms of modal verbs are only in some respects classified as such from the point of view of meaning (Quirk et al. 1985: 220), they rather have a hypothetical meaning which implies present or future time reference.

The directive use of isolated if-clauses in Australian English fits the general pattern exhibited by indirect directives. Ellipsis is identified by Brown/Levinson (1987: 227) as one of the main mechanisms of indirection. In the reconstruction of Stirling’s examples, omitted material would have to do with “either the speaker’s wish for the act or the hearer’s ability/willingness to perform it” (1999: 281). Their communicative meaning seems to be something along the lines ‘I want you to do X, but I don’t necessarily believe that you will do it’. The if-clause allows the possibility of communicating that X may not
be performed. As Stirling explains (1999: 281), this has to do with the assumption that part of the meaning difference between an assertion and an *if*-clause is the difference between ‘I believe that X’ and ‘I believe that there is some possibility that X’. The use of the *if*-clause allows the expression on the part of the speaker that he/she is not assuming the performance of the act he/she is requesting, that the hearer has an option.

With the aim of providing an exhaustive picture of the circumstances under which isolated directive *if*-clauses are used, Stirling also investigated a corpus of Scottish English (the HCRC Map Task Dialogue Database) (*HCRC Map Task Corpus*), in which conversations between speakers were controlled as regards both their eye contact and degree of familiarity. The analysis of these data showed that directive clauses were used with a higher frequency when participants were unknown to each other, that is, that social distance between the speakers could correlate with the use of an isolated *if*-clause.

3. Corpus study: modality and grammatical status of isolated *if*-clauses

3.1. Corpus description and results

With the purpose of investigating the uses and characteristics of directive isolated *if*-clauses in spoken British English I carried out a corpus study. I selected a spoken corpus since, as is well known, changes in language normally appear first in the spoken medium and then move progressively to writing. Furthermore, previous research points to the fact that conditional clauses are more frequent in spoken English (cf. Biber et al. 1999: 820-825) and that this medium is the preferred one for isolated *if*-clauses to be attested (cf. Stirling’s findings in §2.2).

I analysed data extracted from the *DCPSE* (Aarts/Wallis 2006). It is a parsed corpus of spoken English containing more than 400,000 words from the *London-Lund Corpus (LLC)* (Svartvik et al. 1980) and 400,000 words from the *British Component* of the *International*
Corpus of English (ICE-GB). It amounts to 885,000 words of spoken British English covering the period from 1960 to 1990.

The DCPSE yields a total of 3,019 hits for *if*. Repetitions of the conjunction and *ifs* that did not introduce any clause were discarded from the analysis, the resulting figure of *if*-clauses analysed being 2,867. The tokens were then introduced in a Microsoft Access database and classified according to their status as prototypical conditional clauses, conditional clauses whose main clause is supplied by the context, conditional clauses expressing an offer or directive, isolated *if*-clauses (directive or optative) and other types of *if*-clauses. The latter category includes *wh*-clauses, comparative clauses introduced by *as if*, *what if*-clauses, verbless clauses introduced by *if* and indirect conditional clauses. Table 1 shows the classification of *if*-clauses in the corpus according to these categories.

<table>
<thead>
<tr>
<th>Type of Clause</th>
<th>Number of Occurrences</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototypical conditional clauses</td>
<td>1,843</td>
<td>64.28%</td>
</tr>
<tr>
<td>Prototypical conditional clauses with contextually supplied main clause</td>
<td>14</td>
<td>0.48%</td>
</tr>
<tr>
<td>Prototypical conditional clauses with directive meaning</td>
<td>17</td>
<td>0.59%</td>
</tr>
<tr>
<td>Isolated <em>if</em>-clauses</td>
<td>58</td>
<td>2.02%</td>
</tr>
<tr>
<td>Other types of <em>if</em>-clauses</td>
<td>935</td>
<td>32.61%</td>
</tr>
<tr>
<td>Total</td>
<td>2,867</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 1. Distribution of *if*-clauses in the DCPSE.

As discussed in §2.1, prototypical conditional clauses may also convey commands. Examples (15) and (16) below represent cases of clauses where a conditional softens the effect an imperative may have, giving the hearer the freedom to fulfill the order or not. In the first case, participant A in the conversation encourages participant B to hang up his coat, immediately adding a softening *if you’d like to* in order to give the hearer the choice to do or not to do it. In (16), participant A is giving a direction to participant B where the *if*-clause clearly has no conditional meaning, for the presence of the shop is not dependent on the hearer’s walking in that direction.
Beatriz Mato-Míguez

(15) A: Oh hello <,> you found your way <,>  
A: Emma Smith  
A: Do hang your coat up if you 'd like to <,> (DCPSE DI-A14 0007)  
B: Mmm <„>  

(16) A: Well you know Chapel Street  
B: Yeah up at Islington  
A: Yeah <,>  
A: If you go on a bit you come to <,> a corner shop a big which used to be a  
big Lyon 's <,> with a (DCPSE DI-B06 0008)  
A: Oh you don't know oh

Similarly, examples (17), (18) and (19) do not express a conditional relation between the main and the subordinate clause, but rather some sort of request the hearer is expected to accept. In example (17), the main clause we would be most grateful expresses nothing but the feeling of gratitude the speaker would have if the request were granted. A similar explanation holds for (18) with the main clause I’ll be grateful. In other words, these complete conditional clauses seem to establish a causal relation between the two propositions, one providing the reason for the other.

(17) B: so if we could borrow these and have a look through them and uh then let you have them back we would be most grateful <,> (DCPSE DL-B14 0098)

(18) B: I I’ll be grateful if you lay it on when I do (DCPSE DI A10 0023)

(19) B: and uh if you can if you can uh uh write to us and ask us for some help we 'd be vEry [sic] glad to give it to you (DCPSE DL-A01 0317)

As commented on in §2, isolated if-clauses are formally marked in that they lack a main clause. However, not all occurrences of if-clauses without a main clause can be classified into this group. Some examples of if-clauses without a main clause simply lack it because it is clearly recoverable from the context or its presence would prove redundant in the course of a conversation. They usually constitute a response. An example of this type follows under (20), in which we can easily provide the conditional clause with a main clause along the following lines: “you can still rise above it if you keep it up”. 

If you would like to lead: directive isolated if-clauses in spoken BrE

(20) A: Yes exactly <,>
B: But it shows though you can still <,> rise above it
B: try <,> if you keep it up <,> (DCPSE DI B36 0312)
B: It’s like <,,> It’s like it’s
B: I
A: You can
A: Of course you can if you can place it

As seen in Table 1, the corpus analysis yields 58 examples which can be considered isolated if-clauses. This category contains only six examples of the optative type (cf. example (2) above) and 52 of the directive type. In what follows, the focus is on directive isolated if-clauses, and their modality and grammatical status will be considered in light of several pragmatic and syntactic tests.

3.2. Verb tense and modality of directive isolated if-clauses

One of the formal features associated with the study of directive isolated if-clauses relates to the morphology of the verb group, whether present, past or comprising a modal. It is important to notice, however, that in conditional clauses verbs are usually backshifted. This means that the past tense form of the verb is used for present and future time reference, its meaning being counterfactual (cf. Quirk et al. 1985: 1010). This use of the hypothetical is also found isolated if-clauses, an issue that deserves future research. There is just one such example in my data, shown in (21). Present tense forms are far more common, as exemplified in (22):

(21) C: If you gave it to her Dad (DCPSE: DI-B03)
(22) A: yeah if you wish to wash hands et cEtera <,,> alOng the cOrridor on the extreme rIght- hand right-hand side (DCPSE DL-A07 0513)

However, the majority of examples of directive isolated if-clauses in the corpus (59.6%) show a modal verb, most commonly in the past. Quirk et al. (1985: 219) define modality as “the manner in which the meaning of a clause is qualified so as to reflect the speaker’s judgement of the likelihood of the proposition it expresses being true”. Nine central modal auxiliary verbs used to express modality are distin-
guished both by Quirk et al. (1985: 137) and by Biber et al. (1999: 483), namely can, could, may, might, shall, should, will, would and must. Following their classification, from a semantic point of view, modal verbs can be classified into three categories according to the modal meanings they express:

1. volition/prediction: will, would, shall, should.
2. permission/possibility/ability: can, could, may, might.
3. obligation/necessity: must.

The meanings of permission, obligation and volition involve some kind of human control over the actions and events, while the meanings of possibility, necessity and prediction usually involve human judgement, rather than human control, of the events which are or are not likely to occur (cf. Quirk et al. 1985: 219).

Following the semantic classification mentioned above, the occurrences of modal verbs in isolated if-clauses in the corpus are distributed as shown in Table 2.

<table>
<thead>
<tr>
<th>MODAL</th>
<th>NUMBER OF OCCURRENCES</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Will</td>
<td>2</td>
<td>6.45%</td>
</tr>
<tr>
<td>Would</td>
<td>15</td>
<td>48.39%</td>
</tr>
<tr>
<td>Permission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can</td>
<td>6</td>
<td>19.34%</td>
</tr>
<tr>
<td>Could</td>
<td>8</td>
<td>25.81%</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2. Distribution of modals according to their semantic value in directive isolated if-clauses in the corpus.

As the data in this table show, more than half of the instances include a modal verb expressing volition, commonly in the past, which emphasises the hypotheticality of the illocutionary act in question. This is not surprising since modals are often associated with particular pragmatic uses, e.g., requests or offers, where the past forms tend to have implications of politeness (cf. Quirk et al. 1985: 220). Examples (23) and (24) illustrate the uses of will and would. It is worth noting that almost half of the examples including would show the softening expres-
sion ‘d like to’, which helps to further emphasise the freedom of choice on the part of the hearer, as shown in (25) and (26) by means of two different isolated if-clauses.

(23) A: So if you if you will go downstairs and then you could look through these two postgraduate guides (DCPSE DI-A08 0135)
(24) If anyone would like some ice cream <unclear-words> (DCPSE DI-B58 0285)
(25) A: Now if you ‘d like to put on you your helmet (DCPSE DI-F22 0017)
(26) A: If you ‘d like to engage neutral <,,> (DCPSE DI-F22 0076)

As regards modal verbs expressing permission, they are almost equally distributed among present and past modals. All the examples that follow express a request on the part of the speaker.

(27) A: uhm uhm if you can make it about three (DCPSE DL-C03 0445)
(28) B: well if you can really very handsomely lend me your <,,> car on Wednesday that’s me for Wednesday really because I shall go out to Stadlowe and get back a bit late to really go to anything <,> (DCPSE DL-B10 0619)
(29) A: and uh if you if you could show them there’s enough in what you’ve studied within your course that would relate to literature and maths and and whatever ‘s taught you might just squeak in (DCPSE DI-A06 0187)
(30) B: But if we could just for a moment concentrate on the latter years of the nineteenth century (DCPSE DI-D12 0025)

In addition to examples containing a modal verb, the corpora also yielded several instances with want to. Such examples have been classified separately since they cannot be said to show a modal verb proper, although the use of want to/wanna as a marker of modality is becoming more common in contemporary language (cf. Biber et al. 1985: 484). As Krug indicates (2000: 140), in a very large majority of cases a “monoclausal modal analysis is superior to a biclausal pur-

posive reading (in order) to”. Example (31) is illustrative of this:

(31) If you want to take it (DCPSE DI-B18 0195)

3.3. Grammatical status of directive isolated if-clauses

A particular interesting issue in the study of directive isolated if-clauses concerns their grammatical status. In this section, several
grammatical tests will be applied in order to classify them as incomplete utterances, elliptical sentences or a minor sentence type.

The fact that isolated if-clauses lack a main clause may lead us to consider them as incomplete utterances or elliptical from full conditional clauses. Incomplete utterances are described by Matthews (1981: 40-42) as utterances in which the speaker stops not because he/she expects the hearer to supply the missing words, but for circumstantial reasons, such as the speaker’s reformulation of the utterance in progress, or the interruption by another discourse participant. The resulting fragments have no standing of their own. For Matthews, such utterances are of “no concern to syntax, except a source of confusion in our data” (1981: 41). Examples classified as isolated if-clauses in this study were those that clearly intended to be complete.

Similarly, we cannot consider isolated if-clauses as elliptical for prototypical subordinate clauses. Ellipsis is here understood along the lines expounded by Quirk et al. (1985), Huddleston/Pullum et al. (2002) and other standard reference grammars, that is, as “a reduction through omission” (Quirk et al. 1985: 889) manifested in the following characteristics conveniently summarised by Quirk et al. (1985: 888):

1. The ellipted words are precisely recoverable.
2. The elliptical construction is grammatically ‘defective’.
3. The insertion of the missing words results in a grammatical sentence with the same meaning as the elliptical sentence.
4. The missing expression is recoverable from the context.
5. The missing words are present in the text in exactly the same form.

If we consider them separately, criteria 1, 4 and 5 cannot be applied to isolated if-clauses. Let us consider example (32) from our corpus:

(32) A: Now before we start the engines up I would like you all to go to your machines <,> and we will wheel them around the course so that you get the feel of them <,.>
A: Right
A: We ’re going to do this one at a time <,>
A: If you ’d like to go to your machines (DCPSE DI-F22 0017)
We find here three different instances of directive clauses. If we have a look at the context, it can be argued that there is no element that could constitute their main clause, since they are performed as independent speech acts. Moreover, considering criterion 3, given that we cannot retrieve any material from the context to fill the main clause slot, we cannot know whether its insertion would result in a grammatical sentence with the same meaning and implications. Rather, we have to assume that the presupposed main clause would go along the lines of the speaker’s gratitude or wish for the act requested, as seen in examples (17)-(19) above with the use of prototypical conditional clauses for directive purposes.

As regards criterion 2 about the grammaticality of the clause, isolated if-clauses cannot be considered as defective for they are treated by speakers as complete, functional clauses. In the case of directives, they have the illocutionary force of an indirect request and they are likely to be understood and receive a response on the part of the hearer. As we can see from our example (30), the requests made by means of isolated if-clauses attract the expected response since the speaker goes on giving more orders and expressing his/her satisfaction with the hearers’ acts, as expressed by that’s it.

Another piece of evidence supporting the status of these clauses as non-subordinate is reflected in examples (33)-(37) below, where isolated if-directives function as independent clauses taking dependent clauses, a causal clause in (33)-(36) and a temporal clause in the case
of (37). There is no material in the context that can be considered as the main clause of these *if*-clauses and yet they are not considered as ambiguous by the speakers who attach a subordinate clause to them. Example (37), for instance, can be paraphrased by means of the imperative clause “Hold on just half a minute while I put these potatoes out”.

(33)  B: well if you can really very handsomely lend me your <,,> car on Wednesday that’s me for Wednesday really because I shall go out to Stadlowe and get back a bit late to really go to anything <,> (DCPSE DL-B10 0619)

(34)  B: and if you can get here during daylight hours because of course then the ones that they’re just finishing off they haven’t [sic] got any [sic] electrics [sic] on <several unclear-syllables> (DCPSE DL-C03 0710)

(35)  C: yes <,,> well if you could come to our Office[sic] here because the flats are just (DCPSE DL-C03 0696)

(36)  A: If we go through some very simple biochemistry because as I’m a physicist I’ve got no standing on this (DCPSE DI-F21 0018)

(37)  A: Yes uhm <,,> well if you can hold on just half a minute while I put these potatoes out (DCPSE DI-C05 0174)

Examples (38) and (39) go a step further. In the first case we find a second *if*-clause expressing a condition for the fulfilment of the first one meaning “Draw it on a bit of paper, in case you have a spare piece of paper”. In the second, an initial *if*-clause expresses the uncertainty of the second fact, as dependent on the first, its meaning being “If you happen to come on Tuesday, come to my house and we’ll have a coffee”.

(38)  A: If you draw it on yeah on a bit of paper <,> if you’ve got a spare piece of paper <,> (DCPSE DI-B72 0094)

(39)  C: uhm <,,> if you come over to Tuesday [sic] if you want to come up to my house and have a coffee (DCPSE DL-C04 0229)

Further evidence of the status of isolated *if*-clauses as independent clauses is provided by the fact that they coordinate among them without a main clause, as in (40) below, and also with other clauses which are clearly independent, as exemplified here under (41).

(40)  A: If you’d like to go to your machines (DCPSE DI-F22 0017)
    A: and Gareth if you’d like to lead <,,> (DCPSE DI-F22 0018)
If you would like to lead: *directive isolated if-clauses in spoken BrE*

(41) A: *So if you if you will go downstairs and* then you could look through these two postgraduate guides (*DCPSE DI-A08 0135*)
A: *Uhm and then I’ll show you where the others are*

Finally, isolated *if*-clauses can also be used in combination with imperative clauses giving direct commands in order to soften their imposing nature, as in our example (32) above, part of which is repeated here as (42) for convenience.

(42) A: *Keep* the bike leaning towards you slightly not on your hip but towards you
A: *If you’d like to take a right-hand circle <,> quite tight <,> quite gently* (*DCPSE DI-F22 0021*)
A: *Now use the brake if necessary to stop it <,>*

Examples of directive isolated *if*-clauses are also attested in novels. Consider (43) and (44)\(^2\) below, whose punctuation clearly indicates that they are regarded as independent clauses by the writer. The presence of isolated *if*-clauses in writing seems to suggest that they are, at least, on their way towards conventionalization.

(43) ‘If you could lift up your top’, Mrs Brandon.
(44) ‘Right, *if you’d just all move a bit closer together*’, he bellows. ‘Closer, please!’

4. Isolated *if*-clauses as an example of insubordination

In view of the data presented so far it seems that directive isolated *if*-clauses represent a clear example of what Evans calls “insubordination”, that is, “the conventionalized main clause use of what, on *prima facie* grounds, appear to be formally subordinate clauses” (2007: 367).\(^3\)

\(^2\) Taken from Sophie Kinsella’s (Madeline Wickham’s) *Shopaholic and Baby* (2007: 13), and Jane Costello’s *Bridesmaids* (2008: 168).

\(^3\) *If*-clauses are not the only type of clauses that insubordinate in the history of English. As Traugott (2010: 104) explains, *because*-clauses without main clauses are attested in 17th century plays. In her view, it is likely that spoken
Insubordinated clauses usually look like subordinate clauses, because they show prototypically subordinate characteristics, such as the presence of infinitive, participial or subjunctive inflections on their verbs, subordinate word order, complementizers and so on. However, as they get reanalysed as standard constructions over time, those features “will no longer be restricted to subordinate clauses, so that the term subordinate means, at best, ‘having diachronic origins as a subordinate clause’” (Evans 2007: 370).

Evans distinguishes three functions of insubordinated clauses:

1. Interpersonal coercion, especially commands but also permissives, abilitatives, threats and warnings. According to Evans, the most common type of insubordination is found in various types of clause concerned with interpersonal control, primarily imperatives and their milder forms, such as hints and requests, but also permissives, warnings and threats. All such clauses are face-threatening acts (Brown/Levinson 1987), and insubordinating ellipsis has the effect of putting the face-threatening act off the record. Insubordinated clauses of this type tend to take the form of complements of request, desire or possibility predications; purpose clauses with the implication ‘I say this (in order that X); and conditional clauses with an implicit ‘It would be nice’, etc.

2. Modal framing of various types, including the unattributed evocation of quotation or belief, and other kinds of deontic and evidential use.

3. Marking of various discourse contexts, such as negation, contrastive statements and reiteration, all high in presuppositionality.

Our directive examples clearly fall into the first category. As Evans himself recognises, “the most common function of insubordinated conditionals is to express polite requests” (2007: 380). In his view, the
more conventionalised if-clauses become, the less certain speakers are of what exactly has been omitted.

According to Evans (2007), the diachronic formation of insubordinated clauses takes place in four steps:

1. Stage one. Subordinate construction with explicit main clause. This phase is the normal situation, where a subordinate clause is used as such, as in (45) below from German:

   (45) Ich erinnere nicht, ob sie eine Karte gekauft hatte.
   I remember not, whether she bought.
   ‘I don’t remember whether she bought a ticket’ (Durrel 1997: 378, as quoted in Evans 1999: 371)

2. Stage two. Ellipsis of main clause. In this phase, any grammatically compatible main clause can easily be reconstructed by the hearer. Conversational inference is what determines exactly which main clause is restored. Example (46) from German is illustrative of this.

   (46) [Ich zweifle.] Ob wir richtig sind?
   I doubt if we are right.
   ‘(I doubt) whether we are right’ (Buscha 1976 as quoted in Evans 1999: 372)

   More difficult to test, because of the large number of candidate ellipted clauses, is the question of whether there are any limits on which ellipted clauses can acceptably be reconstituted. When there are significant restrictions on this, as a result of the conventionalised use of the construction, we move to the next stage.

3. Stage three. Conventionalization of ellipsis with certain syntactically permitted reconstructions becoming excluded by convention. There is considerable range in the degree to which restoration of material is conventionalised to a subset of the grammatically tolerated possibilities. In the following case from German, a positive rather than a negative consequence is permitted.
Stage four. Conventionalised main clause use of formally subordinate clause. Grammaticalization of a construction with a specific meaning of its own, where restoration of ellipsed material is not possible. A clear example of this is the concessive use of German wo (‘where’) clauses with subordinate verb order, to which a paraphrase from which they can be derived by deletion cannot be supplied.

The four-stage pathway proposed by Evans, first opens up, then closes, the role of pragmatics. In Evans’ words (2007: 374-375),

first a previously syntacticized subordinate clause, made independent becomes available for pragmatic interpretation; in this phase grammatical formatives get opened up to the pragmatics and become less grammatical. Only in the second phase does degrammaticization occur, as the newly independent clause acquires a more specific constructional meaning.

The isolated if-clauses analysed in this study seem to be placed somewhere between stages 3 and 4 in the process of insubordination, since it has become clear that directive clauses do not need a main clause to constitute an illocutionary act. In addition, if we were to reconstruct their main clauses, not all types of clauses would be allowed.

The process of insubordination is of great interest for theories of historical morphosyntax. The literature on grammaticalization and reanalysis concentrates on diachronic developments in the opposite direction, i.e., the development of subordinate constructions from mate-
If you would like to lead: directive isolated *if*-clauses in spoken BrE

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If you would like to lead: directive isolated *if*-clauses in spoken BrE

It has been asserted that there is a unidirectional pathway from pragmatics to syntax to morphology, one consequence of which is that loose paratactic pragmatic constructions become syntactically as subordinate clauses. Wherever insubordination is situated, it is clear that it goes against the usual direction of change by recruiting main clause structures from subordinate clauses. Insubordination, thus, seems to be a process of degrammaticalization, “a composite change whereby a gramm in a specific context gains in autonomy or substance on more than one linguistic level (semantics, morphology, syntax, or phonology)” (Norde 2009: 120).

Isolated *if*-clauses of the directive type may be seen as insubordinated from either indirect interrogative clauses or conditional clauses, as exemplified here in (49), since both can serve the purpose of issuing an order or request.

(49) (I wonder) *If you can open the door* (it would be nice).

In my view, directive isolated *if*-clauses originate from full conditional clauses for three main reasons. First, optative isolated *if*-clauses have followed a process of insubordination from conditional clauses, which confirms this environment as eligible for insubordination. Consider example (50), where a prototypical conditional clause serves this function of expressing a wish and the consequence it would have if it was fulfilled, and example (51), where an isolated *if*-clause serves the same purpose:

(50) *If only the Spanish would lower their interest rates to weaken the peseta* then Mr Lamont would have more leeway to lower Britain’s without the pound falling out of the Mechanism. (*DCPSE DI-J03 0051*)

(51) uh and uh I mean that is <,> I mean quite aslide from what I I am goin [sic] I 'm now going to Offer [sic] you <,> uh <,> we thought you know *if Only* [sic] we can <,> sort of cast around [sic] and find something [sic] <,> I 'd thought first of all <,> of some kind of resEArch [sic] assistedship in the University of Edinburgh <,> in the University of Birmingham (*DCPSE DL A02 0047*)

Secondly, there is no evidence of indirect interrogative clauses introduced by *whether* being used in isolation. And finally, conditional clauses are a useful device for issuing directives and softening the
face-threatening effect of this type of act (cf. Brown/Levinson 1987) to the point that expressions of the type *if I may* and/or *if you would* have become conventionalised as markers of politeness when trying to avoid a damage to the face of the hearer.

5. Conclusions and paths for further research

The present paper has been devoted to the analysis of the so-called directive isolated *if*-clauses in contemporary spoken British English. Several functions of *if*-clauses in spoken language have been discussed (cf. §2.1.), including the use of conditional clauses to issue offers or requests in the course of a conversation, some of them occurring unconnected to a main clause but yielding the expected response on the hearer. *If*-clauses seem useful for this purpose since they leave the action as contingent on the other part’s preference. The analysis of isolated *if*-clauses in Australian English as reported by Stirling (1999) has also been commented on (cf. §2.2.), describing their uses and features.

Through a corpus study, the modality and grammatical status of isolated *if*-clauses in contemporary spoken British English have been discussed (cf. §3). It has been argued that these clauses cannot be considered as elliptical given that the omitted material is not present in the context (cf. Quirk et al. 1985); rather, they are to be taken as independent functional clauses. Evidence in favour of this interpretation is found in the fact that they are used as stand-alone clauses with subordinate clauses of reason and time specifying them, and also in the fact that they coordinate with prototypical independent clauses. In addition, they alternate with imperative clauses in conversation.

In view of the data analysed in this study, it seems clear that isolated *if*-clauses in English represent an obvious example of insubordinated clauses, that is, the independent use of clauses that seem subordinate in form (cf. §4). The directive isolated *if*-clauses analysed in this study seem to be in the latter stages of this process of insub-
ordination since they do not need a main clause to constitute an illocutionary act. Rather, directive isolated if-clauses seem to be well established in the language and so their meaning is completely conventionalised.

The results found in this study need to be confirmed in future research. An enlargement of the sample and a more exhaustive analysis seem crucial. It seems necessary as well to consider the frequency and uses of these constructions in relation to imperative clauses, with which they seem to share illocutionary force. It is also important to take into account other types of isolated if-clauses, namely optatives, those expressing threats and those expressing surprise, in order to provide a general picture of isolated if-clauses in English.

References


On the automatic analysis of learner corpora. Native Language Identification as experimental testbed of language modeling between surface features and linguistic abstraction

Abstract

Learner corpora as collections of language produced by language learners have been systematically collected since the 90’s, and with readily available collections such as the International Corpus of Learner English (ICLE) (Granger et al. 2009) for English and Falko (Lernerkorpus des Deutschen als Fremdsprache) (Lüdeling et al. 2008) for German there is a growing empirical basis on which theories about second language acquisition and the linguistic system can be informed and applications can be tested.

While most research on learner corpora has analyzed the (co)occurrence of (sequences of) words or manual error annotation, tools for automatically analyzing large corpora in terms of linguistic abstractions such as parts-of-speech, syntactic constituency, or dependency are increasingly available. Similar to the discussion about the role of exemplars vs. prototypes in language, this situation raises the question when to consider surface forms as such and when linguistic categories abstracting and generalizing over surface forms are useful in a corpus-based analysis of learner language.

In this paper, we investigate the issue on the basis of two sets of experiments from the domain of Native Language Identification, the automatic identification of the native language of a non-native writer. The exploration of this experimental testbed is situated in a general discussion of the role and relevance of linguistic annotation of learner corpora.

1 Authors’ email addresses: dm@sfs.uni-tuebingen.de, jkrvinek@sfs.uni-tuebingen.de, sbykh@sfs.uni-tuebingen.de; affiliation: U. of Tübingen.
1. Introduction

To situate the discussion of the automatic analysis of learner data, let us start by looking at the motivations for analyzing learner language. By learner language in the context of this paper we mean the language written by second language learners. While spoken learner language would be equally interesting, written language is more systematically collected and avoids the costly and slow manual transcription process required for spoken language data – though hand-written learner data also requires transcription, and collections of student essays as the most widely available form of learner writing are not necessarily sufficient to reliably interpret learner data, so that depending on the perspective and research question pursued, task-based learner corpora are needed (Ott et al. 2012).

In terms of the different perspectives under which learner data is analyzed, Second Language Acquisition (SLA) research can be regarded as a separate discipline aimed at understanding how languages other than the first language are acquired (Larsen-Freeman 2000). Within this discipline, some researchers view SLA as a part of the general linguistic enterprise aimed at understanding how language as a system works, others focus on the cognitive or on the social factors and emphasize the link to cognitive psychology or sociology. Yet others focus on the effect of different types of interventions and the link to teaching and learning in practice. Regardless of the focus, the analysis of learner data is the essential empirical foundation of SLA research, which is collected in a broad range of contexts: from data obtained in fully controlled experimental settings addressing highly focused research questions, via learner data collected as part of intervention studies in schools, to minimally controlled, general purpose learner corpora collected in international contexts such as International Corpus of Learner English (ICLE) (Granger et al. 2009).
1.1. Learner corpora in SLA

In the more theoretical strands of SLA research, learner data is essentially collected and interpreted to inform the development of new SLA theories or to identify counterexamples to existing theories, parallel to the collection and annotation of native language corpora for the development of linguistics theories in general (cf. Meurers 2005). On the more applied side, the analysis of learner data can also help document and advance our understanding of student abilities and needs, as well as the effectiveness of teaching methods and tools, which are at the heart of Foreign Language Teaching and Learning (FLTL) and related interdisciplinary fields such as Intelligent Computer-Assisted Language Learning (ICALL), where they can in principle be used to inform learner models of intelligent tutoring systems or other tools integrating an automatic analysis of learner language (cf. Meurers 2012).

Considering the research questions for which corpus data is used in SLA, generally speaking they involve observing the occurrence and correlations of linguistic properties. To illustrate this with examples instantiating some of the SLA landscape, consider the early study of Clahsen/Muysken (1986) investigating the acquisition of German word order by native speakers of Romance languages. They assume the following stages of acquisition:

1. S (Aux) V O
2. (AdvP/PP) S (Aux) V O
4. XP V[+fin] S O
5. S V[+fin] (Adv) O
6. dass S O V[+fin]

In the first stage, the learner language is characterized by sentences in which the subject precedes an optionally present auxiliary, followed by the main verb and an object. In the second stage, learners are able to front adverbial phrases, even though they still keep the subject in front of the verb at that stage. And so on. Overall, the stages are characterized by particular sequences of specific categories and functions. The fact that some of these stages of acquisition are well-formed and others are ill-formed in terms of the target language norms is not relevant here. To empirically support or refute such a proposal, one needs to identify learner data instantiating the lexical and syntactic
categories and the grammatical functions used to characterize the stages of acquisition. Corpus data can be relevant here if we can reliably identify the specific corpus instances containing the categories and functions used to characterize the stages of acquisition.

Just as in this early example of SLA research into the stages of acquisition, a broad spectrum of current SLA work crucially relies on empirical learner language patterns, be it for quantitative measures of language development in terms of Complexity, Accuracy and Fluency (e.g., Housen/Kuiken 2009), the identification of characteristic, criterial features of language development (e.g., Hawkins/Buttery 2010), or investigations of indicative underuse or overuse patterns in learner language (e.g., Aarts/Granger 1998; Wiersma et al. 2011; Hirschmann et al. 2013).

For any of those SLA research perspectives, establishing an effective link between the research question and relevant corpus evidence crucially relies on the formulation of the research question in a way that precisely delineates the relevant sets of data. While some research questions involve the occurrence of specific words that can directly be identified using the surface forms in the corpus, SLA research questions often include reference to more abstract linguistic classes such as the grammatical categories and functions in the example above. To be able to efficiently identify subsets of learner corpus data that speak to a given SLA research question, one thus needs a linguistically annotated learner corpus.

1.2. Linguistic corpus annotation

For computational, theoretical, and corpus linguistic research using native language corpora, a wide range of linguistic annotation schemes have been developed (cf., e.g., Garside et al. 1997), ranging from parts-of-speech and morphological schemes of different granularity, via syntactic annotation of constituency or lexical dependency structures, to semantic and pragmatic annotation of word senses, coreference, or information structure.

Such linguistic annotation is fundamentally different from casual notes annotating interesting examples that one happens to come
across, or the typical teacher markup of learner essays in a language-teaching context. Linguistic corpus annotation must provide a reliable index into the classes of data that is of interest, very much like the alphabetic index of a telephone book allows us to reliably access phone numbers by the first letter of the last name.

In consequence, it is essential that an annotation only includes those distinctions, which can reliably be made based on the evidence that is available in the corpus (and its meta-information). Instantiating this point, Marcus et al. (1993, sec. 2.1) discuss the elimination of the nominal adverb tag to avoid inconsistent annotation. This, however, does not mean that annotation schemes with fewer classes necessarily are more reliable. More classes can actually be more reliable if they are more coherent in terms of their observable properties. For example, for the BNC Tag Enhancement Project (Baker 1997), tagging was performed using the more fine-grained CLAWS 7 tagset in order to benefit from a richer inventory of empirically more coherent classes, followed by a conversion back to CLAWS 5 to eliminate class distinctions which for some corpus instances cannot reliably be made. The CLAWS 5 annotation obtained in this indirect way was more accurate than direct tagging with CLAWS 5 would have allowed.

1.3. Linguistic categories for annotating learner language

Returning to the need for linguistically annotated learner corpora to provide access to sets of data which are relevant to SLA research, let us consider which linguistic classes and relations are appropriate for characterizing the interlanguage of language learners and how learner corpora can be reliably annotated to support the identification of those classes.

Consider, for example, one of the most common linguistic corpus annotation tasks, the annotation of parts-of-speech (POS). It is tempting to take one of the readily available POS tagging tools with its standard model trained on native language corpora, which on news corpora are generally reported to achieve above 95% accuracy. While depending on the research question, this can be a reasonable approach, one also needs to be aware of what it amounts to: analyzing learner
language using the native language category system of the target language to be acquired. This conceptual point has important empirical consequences in that the resulting annotation by its very nature will systematically hide those aspects of the learner language that differ from the native language norms.

Let us illustrate this point based on Díaz-Negrillo et al. (2010), who discuss learner data written by Spanish learners of English. Consider the sentences shown in (1).

(1a) RED helped him during he was in the prison.
(1b) you can find a big vary of beautiful beaches

In (1a), the learner used the word during which in terms of its stem is a preposition, yet the distributional slot in which this word is used is that of a conjunction. For the use of vary in (1b) we find a similar mismatch, given that the stem vary is a verb, but in this sentence the distributional slot filled is that of a noun.

As further empirical evidence for determining parts-of-speech, we can also consider aspects of morphology. Consider, for instance, the inflectional suffixes in (2).

(2a) one of the favourite places to visit for many foreigns.
(2b) to be choiced for a job

In (2a), we find the word foreigns, which in terms of its stem clearly is an adjective, but it occurs in the distributional slot of a noun and seems to carry a plural suffix s. In (2b), the use of choiced again includes a distinctive morphological marking, the past-tense suffix -ed, which is in line with the distributional slot it occupies in the sentence. But this contrasts with the fact that lexically the word choice is a noun or an adjective.

In consequence, annotating such learner sentences so that each word is assigned a single POS tag from a standard native tagset would miss the fact that learners do not seem to register and combine distributional, morphological, and lemma properties in the same way as native speakers do. Assigning single native POS tags thus fails to systematically identify characteristic properties of learner language. To develop a POS annotation scheme for learner language that makes it
possible to observe where such mismatches in the empirical evidence for POS arise, one instead needs a tripartite encoding that is closer to the actually observable properties: distribution, morphology, and lexical stem.

Interestingly, the value of identifying such mismatches was independently confirmed by recent SLA research. Zyzik/Azevedo (2009) discuss that second language learners find it difficult to distinguish between word classes among semantically related lexical forms and argue that this may be due to a limited ability to interpret syntactic and morphological cues.

Zooming out to the broader picture, we conclude that annotating learner language with the standard annotation schemes developed for native language can hide important learner language characteristics. In essence, interpreting learner language in terms of native language annotation is an instance of the comparative fallacy, a term introduced by Bley-Vroman (1983: 6) to refer to “the mistake of studying the systematic character of one language by comparing it to another”. In terms of the terminology from computational linguistics, such an interpretation is analyzing a “non-canonical variety” of a language using a “robust” version of the canonical grammar. Under such a perspective, divergences from the native language norm then are classified as errors. Such a perspective seems incompatible with the predominant SLA perspective on learner language as an interlanguage system in its own right. Correspondingly, learner corpora annotated in this way fall short of the target to establish a link between current SLA research questions and the empirical evidence that learner corpora can provide to help answer those questions.

Naturally, the issue of interpreting data against a reference frame that skews or predetermines the outcome is much more general than the question of how to interpret and annotate data in learner corpora. For example, Gil (2001) warns against Eurocentrism in fieldwork and argues for a bottom-up approach to language description. At the same time, he acknowledges that interpretation and systematization of observations is something that cannot be avoided as such – a perspective which seems to be directly in line with our position that linguistic annotation of learner corpora is essential but should be carried out as close as possible to the observed evidence.
Another interesting perspective is added by research in variationist sociolinguistics (e.g., Tagliamonte 2011). Under this research perspective, exploring the choices speakers make in the linguistic system and how those choices are correlated with properties of the speaker and context, it is crucial to precisely define the variation to be studied and when exactly an instance is counted as one of the variants. Bringing the variationist and computational linguistic perspectives together, we think that it is potentially insightful to consider the spectrum between representing variation and providing robust categorization, and to reconsider where linguistic categories come from on that basis. Categories result from generalizations, which require a significant amount of comparable data to be made. So it should not come as a surprise that for the individual interlanguage system of language learners, a language system in flux, it is difficult to identify general characteristic categories.

To inform work on a range of SLA research questions, one will thus need learner corpora which are annotated with rich, multi-layer annotations encoding classes close to the observations as well as more abstract generalizations providing more robust classification to support direct querying of more abstract generalizations. Instead of reusing annotation schemes developed for the annotation of native language corpora to annotate learner corpora, it seems important to always keep in mind the question how fine-grained an annotation scheme must be given a particular research question. The category system used must be sufficiently fine-grained for the variation we want to identify and analyze. Yet, at the same time robustness is needed to ignore other variation in the realization of a category to be identified.

1.4. An experimental testbed for studying linguistic modeling

Complementing the conceptual discussion of the nature and use of linguistic annotation, we would like to develop ways to empirically inform a discussion of the value of linguistic categories in a way that makes it possible to measure the impact of different types of linguistic modeling, from surface observations to more abstract linguistic
characterizations. To find out more about the informativeness of the surface forms and linguistic abstractions, we thus propose to set up classification experiments that make it possible to quantify the impact of the different types of features. For such an investigation to be meaningful, we need a classification task for which a range of linguistic modeling may in principle be relevant. An interesting candidate is the identification of the native language (L1) of the writer of a text written in a second language, so-called Native Language Identification (NLI).

In contrast to early SLA work, which under the Contrastive Analysis Hypothesis compared the native language system with the second language system in general and found that this is insufficient to predict the difficulties arising in SLA (cf., e.g., Ortega 2009, chapter 3), more recent work on L1 Transfer has shifted the focus to the individual learners and their perception of similarities and differences. Transfer as “the influence resulting from similarities and differences between the target language and any other language that has been previously [...] acquired” (Odlin 1989: 27) is known to occur at many levels, including lexical and syntactic aspects as well as more functional aspects of language such as the information structure studied in formal pragmatics (Krifka 2007). NLI thus seems to be well-suited as a classification task to explore the impact of different types of linguistic modeling, from surface to more abstract properties.

In the second part of this paper, we want to showcase this experimental testbed with two strands of experiments. In the first approach, we pursue a data-driven perspective and explore the use of recurring surface strings, to which we then integrate a first step of abstraction using part-of-speech classes. To exemplify the second, theory-driven perspective, we will take a look at a case-study based on syntactic alternations (Levin 1993). Essentially adopting a variationist perspective investigating how the learners make use of the choices offered by the linguistic system, we show that the analysis of syntactic alternations can provide a quantitatively and qualitatively insightful perspective on NLI.
2. Experiments in Native Language Identification

In computational linguistics, NLI is starting to receive some attention (Tetreault et al. 2013) – though with Tomokiyo/Jones (2001), Jarvis et al. (2004), and Koppel et al. (2005) as first NLI publications, it is a very young research focus with fundamental questions still remaining to be addressed. Which language characteristics are relevant, reliable, and sufficiently frequent to be used as features for an NLI classifier? How well can a surface-based approach fare in the task? Which roles can different types of linguistic abstractions play? In the following sections, we investigate these questions to exemplify NLI as a testbed for exploring linguistic modeling between surface-based extensional characterizations and linguistic abstractions.

2.1. Data-driven approach

We start by considering a strictly data-driven approach to NLI as a text-classification problem with the different native languages as the classes. We base our work on recurring n-grams as features, i.e., recurring sequences of words. This perspective builds on the variation n-gram approach to corpus annotation error detection (Dickinson/Meurers 2003; Boyd et al. 2008), which we have further developed and applied to standard NLI datasets (Bykh/Meurers 2012; Bykh et al. 2013).

2.1.1. Data and Tools

We base our study on a subset of the International Corpus of Learner English (ICLE v. 2) (Granger et al. 2009). Following the setup in Bykh/Meurers (2012), we make use of a randomly selected set of essays from seven native languages: Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. We use a data split with 70 essays for training and 25 essays for testing for each of the native languages, so in total we use 490 essays for training and 175 for testing.
the automatic classifier. We only included essays ranging from 500 to 1,000 words.

As preprocessing, we tokenized the essays and removed all punctuation marks, special characters and capitalization. Each essay thus is represented as an array of lower-case words. For part-of-speech tagging, we use the POS-tagger provided by the OpenNLP project.\footnote{See <http://opennlp.apache.org>}

For the machine learning, we employ a linear Support Vector Machine (LIBLINEAR; Fan et al. 2008).

2.1.2. Features

In this first study, we explore three feature types based on recurring n-grams, i.e., sequences of words occurring in at least two different texts of the training set (Bykh/Meurers 2012). We use the word-based n-grams from our previous research as baseline and investigate two new variants of the features: POS generalized (POS-gen) and freely generalized (Free-gen) word-based n-grams. The former feature type incorporates linguistic abstraction of words to parts-of-speech, whereas the latter is based on an abstraction not incorporating linguistic knowledge.

First, we extract all the recurring word-based n-grams with \(2 \leq n \leq 6\) from the training set. Recurring n-grams longer than six were found to be too rare to be effective as features and we do not include unigrams in this experiment because the point here is to compare recurring n-grams with the POS-gen and FREE-gen generalizations, which only come into play for longer n-grams. We start with the bi-grams, which are the same for all three feature types and can serve as a baseline.

Second, based on that extracted feature set, we generate the POS-gen and Free-gen features in the following way: For all word-based n-grams with \(n \geq 3\) we retain the words at the boundaries of the n-grams and replace the rest, i.e., the middle part of the n-grams, by the corresponding POS-tags for the POS-gen features or by a star (*) for the Free-gen features. Figure 1 shows an example for n-grams of length five.
Recurring word-based n-grams: \( He \) gave \( John \) the \( book \)

Recurring POS-gen word-based n-grams: \( He \) VBD NNP DT book

Recurring Free-gen word-based n-grams: \( He \) * * * book

Figure 1. Example for n-grams with \( n = 5 \).

While these three representation options allow us to investigate the impact of a linguistic and a non-linguistic abstraction away from the surface n-gram, one should keep in mind that they are just two of the many abstraction options one could investigate. Also note that we here generalize the word-based recurring n-grams we obtained from the corpus in the first step; another option, to start by collecting all POS-based n-grams recurring in the corpus, is discussed in Bykh/Meurers (2012).

In our previous work, intervals of n-grams always outperformed single \( n \) n-grams, so we here consider only intervals of n-grams, i.e., first the bigrams alone (interval \([2, 2]\)), then the bigrams and trigrams together (interval \([2, 3]\)), etc., up to the interval \([2, 6]\). For the ICLE subset introduced above, we obtain the feature counts for the different intervals shown in Figure 2.

Figure 2. Feature counts.
Depending on the feature type and the interval used, we obtained counts up to 65,000 features. The more abstract the features become, from words-based via POS-gen to Free-gen, the fewer features there are, given that a single more general feature often subsumes multiple specific ones.

For our classification experiments, we use binary feature vectors as input. In other words, each essay is represented by a vector consisting of the values \{0, 1\}. If an n-gram occurs in a particular essay, then the corresponding value in the feature vector for that essay is 1, and 0 otherwise.

2.1.3. Results

Figure 3 presents the classification accuracy for the different \( n \) intervals of n-grams on the held out test data. In the corpus, the seven L1 classes are equally represented, so the random baseline is 14.3%. Using bigrams alone yields an accuracy of 86.9%, which is the same for all three feature types since in the abstractions explored here the n-gram boundaries always consist of words.

![Figure 3. Classification results.](image-url)
The best accuracy for the recurring word-based n-grams as our baseline feature type is 87.4% using the interval [2, 4]. The interesting question now is: what is the effect of the two different types of abstraction, i.e., the linguistic (via POS) and the non-linguistic one (via *)?

As shown in the figure, incorporating non-linguistic abstraction (Free-gen), decreases the classification accuracy. The words in the middle of the n-gram thus include information relevant for distinguishing the different native languages. On the other hand, incorporating some linguistic abstraction using the POS-gen n-grams increased the classification accuracy. The best result for the POS-gen n-grams is 88.6%, for the interval [2, 4], which is the best overall result. In other words, incorporating some linguistic abstraction via POS and including sequences longer than the commonly used trigrams is most successful in providing access to information relevant for NLI.

While a classification accuracy of 88.6% is very promising from the quantitative perspective, the high number and the nature of the features make it difficult to interpret them from a qualitative point of view, so that their value for advancing our conceptual understanding of L1 transfer is limited.

2.2. Theory-driven approach
2.2.1. An alternative perspective

Word-based surface features always encode form and meaning together, and this requires a very high number of features for the approach to be applicable to unseen data, across different topics.

Is it possible to abstract away from the meaning to be expressed to form choices in the linguistic system? If so, that would also make it possible to reduce the number of features required for successful NLI classification across topics. To pursue this perspective, we need to study where the linguistic system provides multiple form options for expressing the same meaning. In essence, this is the method of variationist sociolinguistics (Tagliamonte 2011), but such work has primarily targeted pronunciation variation and lexical choice, whereas here we are dealing with written language and are more interested in
syntactic and morphological variation to avoid lexical meaning and its topic dependence.

A good case study of syntactic variation can be developed on the basis of the seminal work of Levin (1993), who explores the definition of verb classes on the basis of the different syntactic realization options available for realizing a given verb meaning. For example, the verb *give* displays the well-known DATIVE ALTERNATION, which means that the speaker has a choice to realize the objects of *give* in a double-object construction (3a) or as a direct and a prepositional object (3b). Throughout the discussion, we use the alternation names defined in Levin (1993), typesetting them in small capitals.

(3a) He gave John the book.
(3b) He gave the book to John.

Syntactic alternations are rarely discussed in corpus-based second language acquisition research, which may partly be caused by the lack of syntactically annotated learner corpora. A notable exception are Callies/Zaytseva (2011), who used Levin alternations for the classification of learner texts, but they only target a few exemplary syntactic alternations among various other classification features.

### 2.2.2. Syntactic alternations as characteristic features

This section presents a new corpus-based approach to L1-classification of written texts, in which a systematic exploration of the linguistically-motivated syntactic alternations of Levin (1993) plays the role of L1 characteristic features. On the one hand, the study is an example of how linguistic knowledge may contribute to the statistical analysis of corpus data. On the other hand, we will show how computational, data-driven methods may complement and expand linguistic theorizing.

#### 2.2.2.1. Setup

As data for the binary L1-classification experiments discussed in this section we used English essays written by Chinese and English native speakers. The native English essays were taken from the LOCNESS
and the essays of Chinese speakers were taken from the ICLE corpus (Version 2, Granger et al. 2009) of writing by higher intermediate to advanced learners of English. The LOCNESS and the ICLE corpora are comparable in that they consist of texts with similar length and topics written by university undergraduates. We used 720 essays, of which 600 essays were used as a training set and the remaining 120 essays as a held out test set. Both the training set and the test set contain an equal number of the L1 English and L1 Chinese texts.

As regards classification features, we focused on 21 alternations that can be identified quite reliably given syntactic annotation.

2.2.2.2. Identifying alternations

Meurers (2005) and Meurers/Müller (2009) discuss the use of different levels of linguistic annotation for identifying specific linguistic patterns. The present classification task incorporates the search for syntactic patterns, particularly for the verb lemmas and their syntactic arguments. A number of alternations can be identified in a syntactically annotated or parsed corpus.

In (4) we exemplify the AS ALTERNATION. Here press secretary as the second post-verbal argument of appoint can be realized either as an NP (4a) or as an as-PP (4b). Searching for the lemma appoint and two post-verbal NP-arguments will identify sentences such as (4a), whereas the combination of appoint, an NP and a PP headed by the preposition as in a query will yield sentences such as (4b).

(4a) He appointed him press secretary. \( \text{appoint} + \text{NP} + \text{NP} \)
(4b) He appointed him as press secretary. \( \text{appoint} + \text{NP} + \text{PP(\text{as})} \)

Some alternations, including the SIMPLE RECIPROCAL ALTERNATION are more difficult to identify. The verbs displaying the SIMPLE RECIPROCAL ALTERNATION can appear either with or without a PP complement headed by the preposition with, as in example (5). In the absence of the PP complement, the subject must be a collective NP (5b). As illustrated by (5), the prepositional object of agree must be animate. Given that such animacy information is not generally

---

available in syntactically annotated corpora, one cannot rule out misidentifying examples such as (6a), which do not take part in the alternation, as the unavailability of (6b) illustrates.

(5a) Anna agreed with John  NP + agree + animate-PP(with)
(5b) Anna and John agreed.  collective-NP + agree
(6a) Anna agreed with the argument.
(6b) # Anna and the argument agreed.

In our work, we thus focused on 21 alternations which were identifiable with sufficient precision based on the output of parsing the corpus with the BitPar parser (Schmid 2004), trained on the WSJ from the PennTreebank (Marcus et al. 1993).

2.2.2.3. Quantitative Results

The 720 essays used for the experiment were lemmatized using the TreeTagger (Schmid 1994) and syntactically annotated using the BitPar parser. The parser was trained on the enriched WSJ from PennTreebank, where lexical categories contain subcategorization information. This subcategorization information was used in formulating TGrep2 queries (Rhode 2005) to identify the selected syntactic alternations. The 21 binary alternations are represented by 42 features per document, one for each syntactic realization option. The features record the syntactic frame choices in a document: how often do the first and the second variant of a given alternation occur.

For example, the AS ALTERNATION has two possible syntactic frames: ‘verb + NP + NP’ (4a) and ‘verb + NP + PP(as)’ (4b). If a document contains two ‘verb + NP + NP’ and one ‘verb + NP + PP(as)’ instances, the relative frequency of choices for the AS ALTERNATION class is recorded as 2/3 for the ‘verb + NP + NP’ feature and 1/3 for the ‘verb + NP + PP(as)’ feature.

With the setup above, the Weka-SMO classifier (Hall et al. 2009) achieved 63.3% accuracy on the held out test set. This is an improvement over the random baseline of 50%, but it clearly is not realizing the full potential of an analysis of syntactic alternations.

Looking for likely causes for this shortfall, a likely culprit is data sparsity. The average essay length of 790 words is too short to
identify enough instances of syntactic alternation patterns per document. To explore that hypothesis, we created a new corpus by combining each five essays of a given L1 into a new document, which resulted in 120 training and 24 test documents with an average length of 4,000 words. The classification accuracy increased to 70.8%, supporting the assumption that the low number of occurrences of the theory-driven alternation feature is indeed an issue.

2.2.2.4. Adding a data-driven twist

Even increasing the document size does not fully address the underlying problem though. The 21 alternations of Levin (1993) for which we wrote pattern matching rules cover only few English verbs and their possible alternations, and many of these linguistic patterns are quite rare. In addition, the real-life learner data we are analyzing contains variants of the verbs’ argument realizations that differ from the ones characterized by Levin. While the Levin alternations are particularly appealing from a variationist perspective in that the variants in a Levin alternation are known to realize the same meaning (i.e., the same variable), we can relax that requirement a bit and instead pursue a data-driven search for all of the argument realization patterns of a given verb.

The idea is to automatically extract alternations from the given data and then to classify this data by using these data-driven alternations as L1-classification features. To define the data-driven alternations we first record all selection patterns for each verb found in the corpus. We then define data-driven alternation classes, where each class consists of all verbs with the same set of syntactic realization alternatives, which we will refer to as subcat options.

For example, the verb hijack was found with three different subcat options: ‘0’ (intransitive), ‘n’ (NP object) and ‘pn’ (PP and NP object). The verbs smile and infer have the same options as hijack and two additional ones: ‘b’ (SBAR complement) and ‘p’ (PP object). So we assign the verbs smile and infer to one data-driven alternation class, and hijack to another, which is illustrated in (7).
Applying this approach to our corpus, we obtained 846 data-driven alternation classes for a total of 2,610 distinct verbs. The features for the data-driven alternations were defined and computed in the same way as the theory-driven alternations discussed above, with the difference that the data-driven alternations included a feature for each subcat option, whereas the theory-driven approach was based on binary alternations. In total, we obtained 5,287 features for the 846 data-driven alternations.

Table 1 presents the classification results obtained with these automatically extracted data-driven alternations as features, and compares them to the theory-driven implementation of the Levin alternations. The data-driven alternations clearly outperform the theory-driven ones, with 75% accuracy for the actual essay documents and 95.8% for the documents consisting of five essays each.

<table>
<thead>
<tr>
<th></th>
<th>DATA-DRIVEN ALTERNATIONS</th>
<th>THEORY-DRIVEN ALTERNATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 essay</td>
<td>75%</td>
<td>63.3%</td>
</tr>
<tr>
<td>5 essays</td>
<td>95.8%</td>
<td>70.8%</td>
</tr>
</tbody>
</table>

Table 1. Results for data- and theory-driven alternations as L1-classification features.

In the encoding exemplified in (7), which we will refer to as *single-class alternations*, each verb is assigned to a single alternation class, the class offering the full set of subcat options. Thus, even though the verbs *smile* and *infer* also offer the three subcat options of *hijack*, they are not included in the alternation class of *hijack* given that they offer two additional subcat options (‘p’ and ’b’). This differs from the approach of Levin (1993), who defines verb classes in terms of the set of alternations available for that class, and individual verbs can belong to multiple verb classes.

We also explored a second data-driven alternation class encoding, where every verb is included in every alternation class that is characterized by a subset of its subcat options (*multi-class subcat
alternations). In our example (7), the subcat options in (7a) are a subset of the subcat options in (7b), so that in the multi-class approach smile and infer would also be included in the class of hijack. Table 2 summarizes the results for binary classification of Chinese and English texts using single-class and multi-class alternation features. The SMO classifier using the single-class alternation features clearly outperforms multi-class alternation features.

<table>
<thead>
<tr>
<th></th>
<th>SINGLE-CLASS ALTERNATIONS</th>
<th>MULTI-CLASS ALTERNATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 essay</td>
<td>75%</td>
<td>72.5%</td>
</tr>
<tr>
<td>5 essays</td>
<td>95.8%</td>
<td>87.5%</td>
</tr>
</tbody>
</table>

Table 2. Results for different data-driven alternation types as L1-classification features.

2.2.3. Qualitative Results

If language users have a choice in the linguistic system, which option are they going to chose? Does this choice depend on the native language of the speakers? Which alternations turn out to be distinctive for the L1-classification? In order to address such questions, we compare texts from language learners with different native tongues to texts written by native speakers, an approach related to Contrastive Interlanguage Analysis (Granger 1996).

2.2.3.1. Underuse/overuse of patterns

We identify all occurrences of a subcat pattern in the texts of the native English speakers and compare this to the number of occurrences of the pattern in the texts written by the Chinese learners of English to see whether these learners overuse or underuse the pattern compared to the native speakers. Starting the discussion with a pattern that we found to be overused by the learners, the pattern ‘provide NP NP’ displays the verb provide in a double-object construction, as in example (9) from the ICLEv2 corpus.

(9) Universities provides us a chance to live.
Figure 4 shows that only learner documents contained this subcategory option for *provide*. Nine learner texts included one and two documents included two occurrences of the pattern. One could note at this point that it is not surprising that an apparently ungrammatical pattern is only used by learners, but note that nothing in the way we obtained or evaluate the pattern hinges on the question whether a given pattern is grammatical or not, so we can avoid the difficult task of evaluating grammaticality. Here and in the next figures, we report the results for the data set with the 120 large documents (i.e., five essays each). The figures report absolute numbers, which in the actual analysis are normalized by the total number of occurrences of the given verb lemma in the text.

The inverse case, where a pattern is primarily used by the English native speakers, is exemplified by the pattern ‘*see NP as NP*’. An example from the *LOCNESS* corpus is shown in (10).

(10) Now we see it as being absurd in America that women did not have a right to vote.

Figure 5 shows the corresponding distribution of the English and Chinese texts containing the pattern ‘*see NP as NP*’. Only one learner text contained the pattern, whereas it was included in 24 texts written by native speakers.
Indeed, the general pattern ‘V NP as NP’ was also underused by learners, independent of the specific verb, which is illustrated by Figure 6 (showing absolute numbers, which in the actual features were normalized by the total number of verbs in a document).

2.2.3.2. Distinctive alternations

We can identify distinctive pattern-based features by picking out those showing underuse or overuse by the learners. Proceeding to alternation-based features, which kind of alternations are relevant for L1-classification? Here the proportion of the alternation’s variants is what we need to look at. In general, the greater the difference between the distribution of the variants in native and learner texts, the more informative it is for L1-classification.

Consider, for example, the LOCATIVE PREPOSITION DROP ALTERNATION. It is a variation between the ‘V PPloc’ and the ‘V NP’ patterns that is displayed by certain verbs of motion. The pie charts in
Figure 7 show the distribution of the variants in the English texts written by the Chinese and the English native speakers. Each pie chart represents the utterances found with the verbs and patterns belonging to this alternation. Note that the L1 English diagram is smaller than the L1 Chinese one, representing the fact that the English texts contain fewer LOCATIVE PREPOSITION DROP ALTERNATION patterns than the Chinese ones.

![Pie charts showing LOCATIVE PREPOSITION DROP ALTERNATION distribution in English texts.](image)

- V-PPloc (e.g., Martha climbed up the mountain.)
- V-NP (e.g., Martha climbed the mountain.)

Figure 7. LOCATIVE PREPOSITION DROP ALTERNATION distribution in English texts.

Both Chinese and English native speakers prefer to express the goal or the direction of the verb by a direct NP-object: 85% of the Chinese and 63% of the English utterances were expressed by ‘V NP’ pattern. However, the percentage of the texts containing the locative PP argument construction is considerably larger if these texts were produced by L1 English speakers: compare 15% of the Chinese utterances to the 37% of the English ones.

Of course, not all alternations are distinctive L1 features in this way. The popular and frequent DATIVE ALTERNATION, for example, turned out to be indistinctive. As illustrated by Figure 8, the proportion of the ‘V NP NP’ and ‘V NP to NP’ patterns in the Chinese and English documents only differs by 1%.
3. Conclusions

We started our paper by characterizing the role of learner language corpora in SLA and FLTL. On the one hand, the data collected in corpora can provide empirical insights for the development and validation of linguistic theories. On the other hand, the analysis of learner language can also help us document and advance the understanding of student abilities and needs.

We then moved towards the discussion of linguistic annotation of learner corpora, motivated by the need to support effective querying for relevant patterns. For typical SLA research questions, corpus annotation is required to effectively identify the corpus data needed to answer those questions. The issue of linguistic annotation is particularly difficult for the individual interlanguage systems in language development. We illustrated that the standard annotation schemes developed for native language corpora can hide important learner language characteristics. A multi-level annotation with different annotation schemes is needed to provide the right level of abstraction for a range of research questions in a way that avoids a comparative fallacy,
striking a balance between a precise characterization of the data close to the observable properties and the need for robust classification.

Complementing the conceptual discussion of the role of linguistic annotation, we then considered how one can empirically test the role and informativeness of surface forms and linguistic abstractions in the analysis of learner language. For this purpose, we proposed L1 classification as an experimental sandbox for exploring the impact of different features. We presented two strands of experiments: a data-driven and a theory-driven approach. In the data-driven approach we used recurring surface n-gram features for classification of English essays written by learners with seven different native languages and achieved 87.4% accuracy. Our next step was to abstract away from the words within each n-gram feature to words with the same part-of-speech (linguistic generalization) and to any words occurring there (non-linguistic generalization). The comparison of these two abstractions showed that only the linguistically-motivated generalization to POS-classes improved the result.

While the data-driven approach does very well in quantitative terms, the huge feature sets encoding thousands of exemplars make it difficult to qualitatively interpret the features in terms of L1 transfer. In the theory-driven approach, we therefore start from the linguistic modeling. We used syntactic verb alternations and investigated whether those are useful as L1-classification features. While the well-researched notion of verb alternations supports a qualitative analysis, the theoretically established verb alternation patterns rarely occur in the learner data, so that the automatic L1-classification of short texts on that basis cannot reach the level of performance of the surface-based approach. We thus explored the extraction of alternation classes in a data-driven way. The resulting hybrid of a theory-driven perspective with a data-driven extension of its applicability to the real-life data at issue shows promise in providing competitive quantitative performance with qualitatively interpretable features. In our future work we thus intend to pursue this hybrid perspective further, integrating a wider range of linguistic modeling.
References


*ICLE* = Granger et al. (2009).


‘Adjective + whether/if-clause’ constructions in English. An exploratory corpus-based study

Abstract

This paper presents an exploratory corpus-based study of the syntactic constructions associated with the ‘adjective + whether/if-clause’ pattern in English including some lexico-grammatical associations of the pattern such as the lexical sets of adjectives most frequently attracted by it, the negative or non-assertive contexts in which it occurs and the explicit or implicit expression of polarity and alternatives. The aim is two-fold: the first aim is to document the uses of whether/if-clauses in construction with adjectives, which have received comparatively less attention than their use in verbal complementation or as conditional adverbials. The second aim is to portray the ‘adjective + whether/if-clause’ pattern as an illustration of “lexical grammar” (Sinclair 2000), that is, of the inseparability of lexis and grammar. It is suggested that some of the notions developed within current corpus-oriented approaches to account for the lexis-grammar interface (“semantic colligation” (Gabriellatos 2007), ‘collostructional analysis’ (Stefanowitsch/Gries 2003)) should take a step forward in order to accommodate the intricate interactions of grammatical structures, such as the ‘adjective + whether/if-clause’ constructions, with complex combinations of specific lexical classes and more abstract categories such as polarity or negation.

1. Introduction

Subordinate whether/if-clauses have traditionally been discussed in the context of verbal complementation (She asked me whether/if I could...
help her), where their nominal and interrogative nature is usually highlighted, in contrast to the conditional character of their adverbial counterparts (If I was in your place I would agree to it / Whether you’re there or not it gets to you). Most major reference grammars include sections devoted more or less specifically to these two main uses of whether- and if-clauses.²

However, whether/if-clauses share another context which has received comparatively less attention in most reference grammars, both traditional and corpus-based, and to which I will informally refer as an adjectival context, as illustrated by examples like the following:

(1) When Crawford was offered the starring role in a new British stage farce, he was unsure whether to accept. <HRF 1245>
(2) In fact, Dot wasn’t at all sure if the dogs in London had names. <AC5 1335>
(3) He said it was not clear whether the device was intended for an animal or a person. <AK2 753>
(4) Apart from the fact that Suragai had no rank, it was doubtful if he was human. <FSE 3647>

Examples like these are sometimes mentioned by grammars as illustrations of the use of whether/if-clauses as “dependent yes/no interrogative clauses expressing indirect questions” (Biber et al. 1999: 690) or “closed interrogatives” (Huddleston/Pullum et al. 2002: 973), that is, the same kind of subordinate (nominal) clauses that appear as complements or objects of verbs such as ask, wonder, decide, doubt, etc.

The popular ComprehensivE Grammar of the English Language (Quirk et al. 1985), for instance, only gives a couple of examples of the ‘adjective + whether-clause’ pattern in a general section devoted to “Adjective complementation by a wh-clause” (Quirk et al. 1985: 1225-1226) and highlights the fact that most of the adjectives that appear with a following wh-interrogative clause are “intrinsically negative in meaning” or appear in “non-assertive contexts” (Quirk et al. 1985: 1225). If-clauses are not even mentioned in this or actually any other section of Quirk et al. (1985) as possible adjectival complements.

---

although they are discussed – along with whether-clauses – as complements of verbs.

In the volume devoted to *Nouns and Adjectives* (Francis et al. 1998) the only references to the use of whether-clauses in examples like (1) and (3) above are found again in general sections devoted to the “ADJ *wh*” pattern (pp. 402-403) and the “*it* v-link ADJ *wh*” pattern (Francis et al. 1998: 488-490), but, again, no reference is made to if-clauses as possible alternatives to whether-clauses in these adjectival contexts. Interestingly, the authors include a section devoted to the “*it* v-link ADJ *when/if*” pattern (Francis et al. 1998: 491-493), illustrated by examples like *It is horrible when your friends seem to be letting you down* and *It will be amazing if they make another record together*. Consistent with their strict corpus-driven approach, the authors merely take note of the existence of the pattern and list the adjectives most frequently found in it, but no attempt is made to provide a syntactic analysis of the pattern or to determine whether the if-clause that follows the adjective is more like a typical adverbial conditional clause (*If they make another record together, I will definitely buy it*), more like the typical nominal interrogative clause (*I wonder if they will make another record together*) or a different type of if-clause altogether.

A slightly more detailed account of whether/if-clauses in an adjectival context is to be found in Huddleston/Pullum et al. (2002: 975-979), who frame it within their general discussion of “subordinate interrotagatives”. The authors mention some adjectives in the list of words “licensing […] subordinate interrogatives” (*certain, clear, doubtful, questionable*, etc.), a list that includes mainly verbs (*ask, inquire, wonder, decide, doubt*, etc.) and a few nouns (*concern, issue, problem, question*), and briefly illustrate the grammatical constructions where whether/if-clauses are “licensed” by these adjectives.

Thus, our primary goal in studying the patterning of whether/if-clauses with adjectives is to document a set of constructions about which grammars do not say much, assuming that these clauses are essentially the same that appear in verbal contexts, i.e., subordinate nominal interrogative clauses, neatly differentiated from their adverbiaal conditional counterparts. In so doing, we will challenge the traditionally assumed interrogative nature of whether/if-clauses occurring
in adjectival contexts and call for a view that integrates their nominal and adverbial uses.

A second, more general goal is to contribute to a line of research that has been gaining ground within the corpus linguistics approach under different names, including “lexical grammar” (Sinclair 2000), “pattern grammar” (Hunston/Francis 1999), “collostructional analysis” (Stefanowitsch/Gries 2003), and elsewhere in models like Construction Grammar (Goldberg 1995, 2003, 2006) and cognitive linguistics (Langacker 1987, 1991, 2008). From different methodological perspectives, all these approaches are interested in unveiling the deeply-rooted connections between lexis and grammar or, even more radically, the “inseparability of lexis and grammar” as expressed by Schulze/Romer (2009: 1) in their preface to a volume specifically devoted to “exploring the lexis-grammar interface”. In this sense, the different ‘adjective + whether/if-clause’ constructions prove to be perfect illustrations of that inseparability, since the most relevant aspects pertaining to their meaning, function and distribution cannot be solely attributed to lexical or grammatical features, but to the joint interaction of both.

2. Method

Our study is based on the analysis of examples collected from the British National Corpus (BNC) (see Burnard 2000) using the online interface developed by Mark Davies at Brigham Young University (Davies 2004-) and the BNCweb3 interface hosted at Lancaster University (see Hoffmann et al. 2008).

The search for the POS tag “AJ*” immediately followed by subordinator whether in the BNC returns 1,543 hits, which gives a frequency of 15.69 instances per million words; a similar search for the tag “AJ*” immediately followed by if returns 9,289 hits, that is, 94.48 instances per million words. The higher frequency of the ‘adjective +

3 See <http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php>.
The ‘if-clause’ pattern is obviously related to the much more frequent appearance of subordinator if in the corpus (2,575.49 instances per million words) as compared with whether (360.41 instances per million words).

For the exploratory purposes of this paper we restricted our analysis to the 20 most frequent adjectives immediately preceding a whether-clause (1,325 examples), which were also studied in combination with if-clauses (Table 1) and the 20 most frequent adjectives immediately preceding an if-clause (3,241 examples), which were also studied in combination with whether-clauses (Table 2). The total number of examples examined was 4,793, of which 1,354 correspond to the pattern ‘adjective + whether-clause’ and 3,439 to the pattern ‘adjective + if-clause’.

Examples were manually annotated for: a) syntactic construction; b) form and meaning of the adjective, and c) contextual features such as the presence of negative or non-assertive elements and alternative coordination by means of or. All these aspects will be discussed in the following sections of this paper.

<table>
<thead>
<tr>
<th>ADJECTIVE</th>
<th>Whether-CLAUSE</th>
<th>(If-CLAUSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sure</td>
<td>351 (336)</td>
</tr>
<tr>
<td>2</td>
<td>Clear</td>
<td>257 (57)</td>
</tr>
<tr>
<td>3</td>
<td>Doubtful</td>
<td>214 (95)</td>
</tr>
<tr>
<td>4</td>
<td>Uncertain</td>
<td>103 (8)</td>
</tr>
<tr>
<td>5</td>
<td>Unclear</td>
<td>94 (5)</td>
</tr>
<tr>
<td>6</td>
<td>Unsure</td>
<td>70 (20)</td>
</tr>
<tr>
<td>7</td>
<td>Questionable</td>
<td>57 (5)</td>
</tr>
<tr>
<td>8</td>
<td>Certain</td>
<td>51 (10)</td>
</tr>
<tr>
<td>9</td>
<td>Debatable</td>
<td>39 (3)</td>
</tr>
<tr>
<td>10</td>
<td>Immaterial</td>
<td>15 (2)</td>
</tr>
<tr>
<td>11</td>
<td><strong>True</strong></td>
<td><strong>15</strong> (105)**</td>
</tr>
</tbody>
</table>

Tables 1 and 2 show raw frequencies. These raw frequencies were compared with different relevance measures supported by the BNCweb interface (log-likelihood, t-score, z-score, etc.) to check relative levels of attraction between adjectives and whether/if-clauses: since no significant differences were found, raw frequencies were used to select the adjectives included in this study. See Gries (2008, 2010) for the discussion of some of the most widely-used statistical measures of co-occurrence in corpus linguistics.
<table>
<thead>
<tr>
<th>ADJECTIVE</th>
<th>Whether-CLAUSE</th>
<th>(If-CLAUSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Undecided</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>Irrelevant</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>Unknown</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td><strong>Possible</strong></td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>Bothered</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>Arguable</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>Worried</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>Relevant</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>Obvious</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>1,325</strong></td>
</tr>
</tbody>
</table>

Table 1. The 20 most common adjectives immediately preceding *whether*-clause (and comparable frequencies with *if*-clause).

<table>
<thead>
<tr>
<th>ADJECTIVE</th>
<th>If-CLAUSE</th>
<th>(Whether-CLAUSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Better</td>
<td>361</td>
</tr>
<tr>
<td>2</td>
<td><strong>Sure</strong></td>
<td>336</td>
</tr>
<tr>
<td>3</td>
<td>Grateful</td>
<td>319</td>
</tr>
<tr>
<td>4</td>
<td>Surprised</td>
<td>233</td>
</tr>
<tr>
<td>5</td>
<td>Essential</td>
<td>163</td>
</tr>
<tr>
<td>6</td>
<td>Sorry</td>
<td>161</td>
</tr>
<tr>
<td>7</td>
<td>Necessary</td>
<td>153</td>
</tr>
<tr>
<td>8</td>
<td>Useful</td>
<td>151</td>
</tr>
<tr>
<td>9</td>
<td>Nice</td>
<td>145</td>
</tr>
<tr>
<td>10</td>
<td>Good</td>
<td>145</td>
</tr>
<tr>
<td>11</td>
<td>Helpful</td>
<td>140</td>
</tr>
<tr>
<td>12</td>
<td>Important</td>
<td>130</td>
</tr>
<tr>
<td>13</td>
<td>Possible</td>
<td>114</td>
</tr>
<tr>
<td>14</td>
<td>Easier</td>
<td>112</td>
</tr>
<tr>
<td>15</td>
<td>Surprising</td>
<td>107</td>
</tr>
<tr>
<td>16</td>
<td>True</td>
<td>105</td>
</tr>
<tr>
<td>17</td>
<td>Different</td>
<td>98</td>
</tr>
<tr>
<td>18</td>
<td><strong>Doubtful</strong></td>
<td><strong>95</strong></td>
</tr>
<tr>
<td>19</td>
<td>Effective</td>
<td>89</td>
</tr>
<tr>
<td>20</td>
<td>Happy</td>
<td>84</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>3,241</strong></td>
</tr>
</tbody>
</table>

Table 2. The 20 most common adjectives immediately preceding *if*-clause (and comparable frequencies with *whether*-clause).
3. Syntactic constructions associated with the ‘adjective + whether/if-clause’ pattern

A quick look at Tables 1 and 2 reveals an obvious lack of coincidence between the most frequent adjectives immediately followed by a whether-clause and those followed by an if-clause. The analysis of the examples included in our corpus shows that this asymmetry can be related to the syntactic functions of the two types of clauses in an adjectival context. On the one hand, Table 2 shows that the large majority of adjectives most frequently found preceding if-clauses (with the sole exceptions of sure and doubtful) hardly ever occur with whether-clauses. The syntactic analysis reveals that this coincides with a more or less obvious adverbial-conditional use of the if-clause where the adjective and the if-clause are merely adjacent, but they have no direct syntactic relationship with each other, that is, they are not constructionally related (examples (5)-(7)). This use represents 88.73 % of the total number of if-clauses in Table 2.

(5) I thought perhaps things would be better if we had some time away from each other. <HA5 51>

(6) Good information is essential if people are to make informed choices about services. <GXJ 4519>

(7) Things would have been very different if they had stayed put. <CDB 990>

The remaining 11.27 % of if-clauses in Table 2 correspond to their nominal-interrogative uses and they only co-occur with two adjectives, sure\(^5\) and doubtful. Significantly, these are the only two adjectives in that table that attract more whether- than if-clauses.

(8) I was tempted to curtsey in return but I wasn’t sure if my knees were up to it. <A0F 2428>

---

5 Of the 336 cases of if-cl immediately following sure, 269 (80 %) are nominal (example i), whereas the remaining 67 (20 %) are adverbial conditional clauses not syntactically related to the adjective (example ii):

(i) Alexandra was not sure if she should sit down before she was invited to, but by now she hardly cared. <H8X 1641>

(ii) I’m sure if he wants your interference he’ll ask for it. <HA6 317>
It was doubtful if Rose heard the last part of her remark. <GVP 985>

With *whether*-clauses the situation is exactly the opposite, as suggested by Table 1. *Whether*-clauses can also have an adverbial-conditional use (“exhaustive conditional”, according to Huddleston/ Pullum et al. 2002: 973; “alternative conditional-concessive”, according to Quirk et al. 1985: 1099-1101) in an adjectival context, but this adverbial use only represents a meagre 1.81 % (24 examples out of 1,325) and is practically restricted to just two adjectives, *true* and *possible*, in Table 1. Again, it is revealing that these are the only two adjectives in Table 1 with significantly larger frequencies of *if*-clauses following them. With these adjectives the adverbial *whether*-clause specifies the (sometimes conflicting) conditions or alternatives under which the state of affairs reported in the matrix clause is true or possible:

(10) This is true whether you are reading poetry, listening to it or writing it. <F9T 650>
(11) Expansion or reformulation should be possible whether or not the first result is positive or negative. <H0S 243>

The remaining 1,301 examples of *whether*-clauses (98.19 %) reflected in Table 1 correspond to those cases where the *whether*-clause is found in construction with a preceding adjective (their traditional nominal-interrogative use), rather than being merely adjacent to it. This happens with 18 out of the 20 adjectives in Table 1.

(12) It was not clear whether she was yet the French officer’s mistress. <B20 598>
(13) Then he hesitated, uncertain whether to try the cottage, the garage flat or the rear garden. <H8T 226>
(14) It is debatable whether calculating the price is an art or a science. <J6N 699>

Figure 1 summarizes the relative frequencies of the nominal and adverbial uses of *whether*- and *if*-clauses in an adjectival context:
The following sections concentrate on the nominal uses of whether/if-clauses in an adjectival context, that is, those uses where the subordinate clause is found in construction with a preceding adjective. Table 3 shows the raw frequencies of whether/if-clause in the three kinds of construction where they co-occur most frequently with particular adjectives: i) as extraposed subjects (EXT S), ii) as extraposed objects (EXT O), and iii) as complements of the adjective within a complex adjective phrase (ADJ COMP). Figure 2 summarizes the overall distribution of whether- and if-clauses in the three relevant constructions.
3.1. Whether/if-clause as (extraposed) subject [EXT S]

In this construction, the adjective functions as a predicative complement in clauses with attributive or relational verbs (mainly be, but also become, remain, etc.) and the whether/if-clause is an extraposed subject:

(15) It was not altogether clear whether there was a charge of indecent assault. <CS1 1559>
It was **uncertain** whether she was offering him tea or dinner or nothing. <FRH 403>

It remains **unclear** whether the abolition of corporal punishment in schools results from a spread of child-centred views or not. <AN5 1381>

It is **doubtful** if he will remain at the address long enough to accept service of the summons. <EVK 1076>

It’s **questionable** if you could change that perception at a national level. <K5K 1403>

Table 3 and Figure 2 show that this is the most frequent construction associated with *whether*-clauses in an adjectival context, both in terms of the number of adjectives that appear in the construction (14 out of 18) and in terms of raw frequencies (747 occurrences, 57.42 % of all nominal *whether*-clauses). Most adjectives (clear, doubtful, unclear, questionable, debatable, immaterial, etc.) favour clearly this construction over the other two. Figure 2 also shows that *if*-clauses are much less frequently found as extraposed subjects of predicative adjectives; the construction is practically limited to doubtful, which accounts for 91 cases, and clear, with 32 occurrences. Figure 3 summarizes the distribution of *whether/if*-clauses in combination with particular adjectives in the [EXT S] construction. If we consider together all cases of extraposed subjects, 83.84 % correspond to *whether*-clauses and the remaining 16.16 % to *if*-clauses.

The notion of extraposition that is customarily applied to clauses like those highlighted in examples (15)-(19) is intended to account for the
fact that those clauses, which are the logico-semantic subjects of the predicative adjectives that precede them, do not occur in the canonical unmarked position for subjects, that is, initially in the clause and immediately preceding the predicate. Instead, whether/if-clauses tend to appear finally in the clause and immediately following the predicative adjective, being concurrently replaced in the canonical subject position by an expletive or anticipatory it. This is a well-documented textual phenomenon affecting many other types of clauses as a result of the so-called “end-weight” principle (Quirk et al. 1985: 1362; Wasow 2002: 3) that tends to move heavy constituents to the final positions in the sentence. However, from a strict corpus-linguistic point of view, extraposition is not an acceptable notion, especially if interpreted in a dynamic sense, that is, implying some kind of “movement” transformation, as assumed in early transformational-generative accounts (e.g., Rosenbaum 1967). These are explanatory devices suggested by the linguist, not observable evidence and are as such rejected by most corpus linguists (see Francis 1993 or Hunston/Francis 2000: 156-158, amongst others) even if they sometimes keep using terms like extraposition as convenient names to describe observed patterns. However, it is worth noticing that even in a purely static interpretation, the term extraposition tends to suggest that subject whether/if-clauses are located at an extraordinary or unusual position. This may be true when compared with subjects in general, but it is not at all true when clausal subjects are considered, whose dominant, most typical position is precisely that which traditional accounts tend to call extraposed, as proved by Kim (2005) with ample corpus evidence. In the case of subject whether/if-clauses, the post-adjectival position is the unmarked, neutral position, whereas non-extraposed or fronted subject whether-clauses are certainly rare, and if-clauses seem to be excluded from that position altogether, this being one of the differences that most grammars mention when comparing the distribution of both types of clauses (Quirk et al. 1985: 1054; Huddleston/Pullum et al. 2002: 973) and an additional argument to support the marked nature of this position for the clauses under discussion:

(20) Whether it was a social disaster is less certain. <CTW 802>
Whether the Universe will expand forever or whether it will finally contract is not yet clear. <H8K 100>

Whether such pessimism was justified was hardly relevant. <GW3 1088>

### 3.2. Whether/if-clause as (extraposed) object [EXT O]

A very marginal use of both whether- and particularly if-clauses in an adjectival context is as extraposed objects in the so-called complex-transitive construction, with the adjective again functioning as a predicative complement. In our corpus there are only 26 examples of extraposed object whether-clauses (2% of all nominal whether-clauses), which co-occur with the adjectives clear (19 cases), unclear (4), uncertain (2) and doubtful (1), and just 3 examples of extraposed object if-clauses (0.64% of all nominal if-clauses), all of them with the adjective clear.

The construction may have two variants.

- With anticipatory it: typically with the verb make and occasionally with leave. Examples:

  (23) You should make it clear whether you will phone back or whether you wish to be telephoned. <EEB 1100>

  (24) Please make it very clear if Christian Aid owes you money, preferably by attaching a yellow to the account. <HPH 361>

  (25) Dunning (1901) left it unclear whether fascist demonstrations could be construed as a genuine attempt to convert people to a point of view or to provoke by insult. <CS6 1122>

- Without anticipatory it: in this case the verb and the adjective form a kind of verbal periphrasis and the construction is restricted to the combination make clear; in any case, the whether/if-clause still behaves as an object in relation to the adjectival predicate. Examples:

  (26) Neither the U.S. nor Russian statements made clear whether the crash occurred on or beneath the surface. <K5M 8952>

  (27) The lease should make clear whether windows are included in or excluded from the demise. <J6R 220>
The argument against extraposition stated above in relation to subject clauses is even stronger in the case of these object clauses without anticipatory *it* since there are no object *whether/if*-clauses that occur immediately after the verb, thus facilitating the formation of verbal periphrases like *make clear.*

Object *whether*-clauses can become (extraposed) subjects when the main verb is passivized:

(29) It was not made clear whether DEA agents would be ordered out of Mexico. <HLL 663>
(30) It should be made clear whether the stated price is inclusive of meals. <EA9 3010>

3.3. *Whether/if*-clause as complement of the adjective [ADJ COMP]

A third, rather frequent use of *whether/if*-clauses is as complements of the adjective within an adjective phrase (AP). This AP typically functions either as a predicative complement in clauses with relational or attributive verbs, mostly *be*, but also *seem, remain* or *feel* (examples (31)-(35)), or as a kind of “supplementive adjective clause” (Quirk et al. 1985: 424-426, Valera-Hernández/Rizo-Rodríguez 1998) attached mainly to the subject of the superordinate clause (examples (36)-(38)). In both cases, the AP is typically predicated of a human subject since, as we will see in §4.1 below, the adjectives that appear in this construction refer to properties that can only be predicated of humans, such as uncertainty, concern, etc.

(31) She was never quite sure whether it was Lacan’s symbolic phallus he was referring to or his own real one. <ANY 741>
(32) Gubbins seemed uncertain whether he could pull it off. <GSX 168>

6 Cf. similar periphrases of ‘*make* + adjective’ that take *that*-clauses, rather than or in addition to *whether/if*-clauses, as object: *make sure, make clear, make certain, make plain, make explicit*, etc.
(33) The markets still remain unsure whether interest rates are rising to keep the economy in check. <A5G 120>

(34) In fact, Dot wasn’t at all sure if the dogs in London had names. <AC5 1335>

(35) But as long the rent of £10,000 a year continues to be paid, council officers aren’t bothered if Mrs Maxwell lives there or not. <K1E 1504>

(36) Uncertain whether to be flattered or insulted, Margaret smiled primly. <BP7 813>

(37) Bourne had come out, unsure whether anything had happened or not. <GW3 3379>

(38) He looks up, uncertain if he’s being taken seriously. <HH3 12133>

As shown in Figure 2, the [ADJ COMP] construction is slightly less frequent than the [EXT S] construction if whether- and if-clauses are considered together (47.96 % vs. 50.4 %), but it is the most frequent construction for nominal if-clauses in an adjectival context (68.52 %) as a result of their very frequent use as complements of the adjective sure: 269 out of 320 occurrences of if-clauses in the [ADJ COMP] construction appear as complement of this adjective (see Figure 4).

The [ADJ COMP] construction is in fact much more restricted than the [EXT S] construction in terms of the adjectives that appear in it. The adjective sure alone accounts for 73.11 % of all whether/if-clauses in the [ADJ COMP] construction, followed at a very long distance by unsure (10.61 %), uncertain (6.25 %) and certain (3.89 %); a few other adjectives — bothered, doubtful, worried, unclear, clear — appear only marginally in the construction. In any case, as in the other two constructions discussed above, whether-clauses appear much more frequently than if-clauses as complements of all the adjectives present in the construction (Figure 4), which results in a global figure of 62.26 % of whether-clauses vs. 37.74 % of if-clauses.
Connected with the discussion of extraposition above, it is worth noticing that when *whether*-clauses function as complements of an adjective they are occasionally extraposed, although in this case extraposition takes place in the opposite direction, i.e., the clause is fronted to initial position:

(39) Whether that’s er, I’m not too sure. <HE2 58>

(40) Whether they are still critics I am not sure. <FSV 311>

4. Lexico-grammatical associations of the [EXT S/O] and the [ADJ COMP] constructions

In the preceding section we have identified the main syntactic constructions where an adjective is constructionally related to a *whether/if*-clause, namely the [EXT S/O] constructions, where the adjective is a predicative complement of an extraposed subject or object *whether/if*-clause, and the [ADJ COMP] construction, where the *whether/if*-clause is a complement of the adjective within the structure of an adjective phrase.
As many corpus-based studies have repeatedly shown over the last two decades, most linguistic features, whether lexical or grammatical, do not occur alone, but tend to be found “in association with other linguistic and non-linguistic features” (Biber et al. 1998: 5). In this section, the [EXT S/O] and the [ADJ COMP] constructions are explored in terms of some lexico-grammatical features associated with them. The following three aspects will be considered: (i) semantic colligation, i.e., lexical sets of adjectives most-frequently co-selected with the [EXT S/O] and the [ADJ COMP] constructions (§4.1.); (ii) non-assertiveness (§4.2.), and (iii) polarity and the expression of alternatives (§4.3.).

4.1. Semantic colligation

The term “semantic colligation” was coined by Gabriellatos (2007) as “a hybrid between colligation and semantic preference”, colligation being traditionally understood as the co-occurrence of lexis and grammatical categories (Firth 1957: 181; Stubbs 2001: 65) and semantic preference as the attraction “between a lemma or word form and a set of semantically related words” (Stubbs 2001: 65). Hence, semantic colligation is something akin to colligational semantic preference, or, more explicitly, “the mutual attraction holding between a sentence class, […] and a set of semantically related words”.7

Looking at the list of the adjectives most frequently found in the [EXT S/O] and the [ADJ COMP] constructions (Table 3), it can be clearly seen that whether/if-clauses tend to co-occur with very specific

---

7 See Stefanowitsch/Gries (2003) for the related notion of “colostruction”, a blend of collocation and construction, defined as the combination of a particular construction and the lexeme or lexemes that are attracted to it. “Collostructional analysis” is intended as “an objective way of identifying the meaning of a grammatical construction and determining the degree to which particular slots in it prefer or are restricted to a particular set or semantic class of lexical items” (Stefanowitsch/Gries 2003: 215). See also Gries/Stefanowitsch (2004).
sets of adjectives that can be grouped under the following broad semantic categories:8

- **(UN)CERTAINTY**: certain, clear, obvious, sure, uncertain, unclear, undecided, unknown, unsure [apparent]
- **OPINION**: arguable, debatable, doubtful, questionable [moot, arbitrary, convinced]
- **(IR)RELEVANCE**: immaterial, irrelevant, relevant, [unimportant]
- **CONCERN**: bothered, worried [concerned, unconcerned, indifferent, interested]

The two dominant sets, both in terms of the number of types and the number of tokens that they provide, are those we have termed (UN)-CERTAINTY and OPINION. The two sets are actually very close semantically, so that sometimes assigning particular adjectives to one category or another is itself debatable. In any case, considered together these two sets account for 96.83 % of all the examples in our corpus where an adjective is found in construction with a whether/if-clause.

The other two sets of adjectives that co-select whether/if-clauses, those expressing (IR)RELEVANCE and CONCERN, are certainly less significant from a purely quantitative point of view (1.64 % and 1.53 %, respectively), but they must also be taken into account in the semantic characterization of the [EXT S/O] and [ADJ COMP] constructions that we shall attempt in the final section of this paper.

(41) It is **immaterial** whether the act occurs in private or in public. <HXE 2142>

(42) It’s totally **irrelevant** if I was married. <CD6 1199>

(43) Promoters, bookers and pub landlords aren’t **bothered** whether you are going to be the next U2. <A6A 88>

(44) I’m not particularly **worried** if players don’t score. <CBG 293>

The obvious semantic connection of the adjectives expressing (UN)-CERTAINTY and OPINION with unsettled matters or gaps of information is what has traditionally prompted the description of whether/if-clause.

---

8 Additional adjectives not included among the 20 most frequent are shown in square brackets.
if-clauses that appear in the context of these adjectives (or related verbs such as debate, doubt, question, etc.) as subordinate interrogative clauses, assuming that in examples like the following there is some kind of “embedded”, underlying or implicit question (see, for instance, Huddleston/Pullum et al. 2002: 972):

(45) It is not clear whether these are true sampling errors. <FP4 1071>  
[Implicit question: ‘Are these true sampling errors?’]

(46) She wasn’t at all sure if Terry was being serious or not. <HGY 790>  
[Implicit question: ‘Was Terry being serious or not?’]

(47) It is also questionable whether it is wise to kill adult males indiscriminately.  
<J2N 212>  
[Implicit question: ‘Is it wise to kill adult males indiscriminately?’]

(48) But it is doubtful if the problem can or should be solved with spending cuts alone. <K59 1338>  
[Implicit question: ‘Can or should the problem be solved with spending cuts alone?’]

However, it is important to realize that the interrogative meaning that can be perceived in these examples is the result of the interaction between the whether/if-clauses, the semantic load of the adjectives attracted by them in the [EXT S/O] and [ADJ COMP] constructions and additional lexico-grammatical features that will be discussed in the next sections, most especially negation. In other words, rather than being inherently interrogative, whether/if-clauses only become semantically interrogative when they are co-selected with and grammatically related to adjectives expressing uncertainty, opinion and similar notions, as in (45)-(48), but not, for instance, when they are co-selected with adjectives expressing (ir)relevance or concern, as in (41)-(44). In this connection, it is relevant to point out that, when that-clauses co-occur with adjectives expressing uncertainty or opinion they assume a similar interrogative sense. To the best of our knowledge, despite their use in examples like the following, that-clauses have never been described as interrogative:

(49) It is not clear that producers of genetically modified organisms will be fully liable for any damage they may cause. <AAP 51>  
[Implicit question: ‘Will producers of genetically modified organisms be fully liable for any damage they may cause?’]
I am not sure that she even knows the word ‘special’. [Implicit question: ‘Does she even know the word ‘special’?']

It is questionable that texts can always be read as favourably as she seeks to do. [Implicit question: ‘Can texts always be read as favourably as she seeks to do?’]

It is doubtful that any words can express what he did to that young girl. [Implicit question: ‘Can any words express what he did to that young girl?’]

Conversely, although some adjectives may be said to be inherently interrogative – doubtful, questionable –, many others – certain, clear, obvious, sure – only assume interrogative connotations under certain circumstances, most typically when they are used in negative contexts and in construction with whether/if-clauses (cf. I’m sure that... vs I’m not sure whether/if...; It is clear that... vs It is not clear whether/ if...).

4.2. Non-assertiveness

A very distinctive feature of the distribution of the [EXT S/O] and the [ADJ COMP] constructions is their systematic occurrence in non-assertive and, more specifically, negative contexts. This negativity or non-assertiveness has in some cases a purely lexical basis (the lexical content of the adjective is inherently non-assertive or negative), but in most cases it is the result of the interaction between lexical and morphological or syntactic factors. Consequently, the following three categories can be distinguished depending on how contextual non-assertiveness is instantiated:

---

9 Quirk et al. (1985: 1225) notice that most of the adjectives that appear in these constructions are “intrinsically negative in meaning” but they relate the presence of subordinate wh-clauses (not only whether-clauses, but also those introduced by interrogative what, when, how, etc.) to “non-assertive” contexts, even though most of the examples they mention are clearly negative and only one is interrogative). Our corpus data reveal that adjectives tend to select whether- or if-clauses preferably in negative contexts, whereas in other non-assertive contexts, such as interrogative, they tend to prefer that-clauses.
• Lexically: the lexical content of the adjective is itself non-assertive, assigning the adjective to the semantic category of OPINION (see §4.1.), and more specifically, to the category of ‘unsettled matters’. Most of these inherently non-assertive adjectives – *doubtful*, *questionable*, *debatable* – are derived from nouns or verbs expressing similar notions:

(53) At this late stage they were **doubtful** whether the National Executive would agree. <HP2 1540>

(54) It was **doubtful** if those two men could have done much worse than he was doing now! <JXV 715>

(55) It’s **questionable** if you could change that perception at a national level. <K5J 1403>

(56) It’s **debatable** whether that statement is true or not but it was certainly pretty blatant. <A6E 614>

• Morpholexically: adjectives marked by a negative prefix, typically [UN-] or [IN-] (in any of its allomorphic realizations (*in-, im-, ir-*); however, negative marking is not enough since not all negative adjectives occur in the pattern: the lexical content of the adjective must be related to any of the broad semantic categories established in 4.1. (CERTAINTY, OPINION, RELEVANCE, CONCERN): *uncertain, unclear, unsure, undecided, unknown, unconcerned, unimportant, immaterial, irrelevant*:

(57) They were also **uncertain** whether it would work. <ACS 1296>

(58) It is **unclear** whether this theory is serious or a misunderstanding. <B7K 1065>

(59) She swallowed hard, **unsure** if she had the nerve to go ahead. <JY3 1562>

(60) It’s **irrelevant** whether he lived or died: the point is that the myth continues. <HH3 11914>

(61) It is wholly **immaterial** whether he has read the document or not. <J7C 648>

• **Lexico-grammatically**: the lexical content of the adjective interacts with the presence of clause negation by means of *not*, or with the presence of semantically negative or non-assertive words in the clause, including adverbs such as *never, hardly, no longer*, etc., pronouns like *nobody, neither*, or expressions like *far from, by no means, without*, or *hard to*:
A report in the weekly Moscow News did not make clear if the insurgents were Armenians or Azerbaijanis. <A2M 258>

You are not bothered whether the house is detached or semi-detached. <K94 438>

Wilkinson is not worried whether he passes his 24-goal record in his first season at Watford. <K4T 1037>

It isn’t relevant whether the creators live or die at the end of ‘The Food of the Gods’. <KAY 701>

It is far from obvious whether terms are being inverted or put in direct order. <H99 1794>

It has never been clear whether he jumped or was pushed. <ACN 2460>

I’m never too sure whether you’re going to be up or not. <G0F 1163>

The system will arouse doubt and uncertainty since no-one will be sure whether he or she can qualify. <AAC 698>

Neither was sure if one of us would be going home or if we’d be split up and put with other hostages as had happened before with the Yanks and the Frenchmen. <FS0 1605>

It is by no means clear whether the omission/commission dichotomy which applies to the crime of murder applies to suicide. <ASK 1620>

The instantiation of a negative or non-assertive context by lexico-grammatical means is particularly relevant, both from a descriptive and a theoretical point of view, as it provides clear evidence of the complex interactions between lexicon and grammar or, more radically, of the impossibility of separating both (Schulze/Romer 2009: 1). The presence of whether/if-clauses in examples like those mentioned above cannot be solely attributed to either the presence of the relevant lexical material, i.e., the adjectives sure, clear, relevant, etc. or to clause negation alone, but to the joint interaction of both. This kind of interactions have been receiving increasing attention within corpus-oriented approaches, like Sinclair’s (2000) “lexical grammar” or Stefanowitsch/Gries’s (2003) “collostructional analysis”. In this connection, it seems relevant to notice that Sinclair (2000: 197-199) explicitly mentions negation (or, more generally, polarity) as one of the many “cross-border categories […] straddling the borders between grammar and semantics”, other such categories being number, possession, tense, modality or voice. Gabriellatos’ notion of “semantic colligation” (Gabriellatos 2007; see §4.1 above) and in particular his “modal colligation” also represents a very promising attempt at explaining the mutual attraction between a specific syntactic class – if-
conditional clauses – and the abstract semantic category of modality. In our case, we have shown the mutual attraction holding in the [EXT S/O] and the [ADJ COMP] constructions between whether/if-clauses and abstract semantic categories like UNCERTAINTY, IRRELEVANCE or UNCONCERN which are not necessarily linked to individual lexical items but which result from the complex interactions between morphological, lexical and syntactic features.

4.3. ‘Or (not)’

One final aspect that we will consider briefly is the connection between the [EXT S/O] and the [ADJ COMP] constructions and the expression of polarity and alternatives. These two notions are often used by grammars in the description of interrogative clauses, so that ‘polar’ questions (also known as ‘yes-no questions’) are “those that expect affirmation or negation” (Quirk et al. 1985: 806), whereas alternative questions “expect as the reply one of two or more options presented in the question” (Quirk et al. 1985: 806). In this context, as we saw in 4.1., whether/if-clauses are typically treated as subordinate interrogative clauses that can be both polar (i.e., ‘yes-no’), as in (72)-(73), and alternative, as in (74)-(75): in fact, polar interrogatives can be interpreted as a special case of alternative interrogatives where the two options are ‘yes’ and ‘no’:

(72) He had not been clear whether or not he should attend the wedding ceremony. <HHC 1671>
(73) Not entirely sure if she was being teased or not, she gave a weak smile. <HGY 891>
(74) It is immaterial whether the time of payment or the time of delivery, or both, be postponed. <HH7 547>
(75) Gabriel sat by his bedside, uncertain if she wished him to live or die. <K8V 3907>

Despite the apparent parallelism between whether- and if-clauses that these examples might suggest, the two types of clauses show a distinct behaviour vis-à-vis the expression of polarity and alternatives.
Thus, in the first place, whether-clauses are much more frequently associated with the expression of explicit alternatives than if-clauses: 44.75% of all whether-clauses in the [EXT S/O] and the [ADJ COMP] constructions include the use of correlative coordinated expressions such as ‘whether ... or whether...’ or ‘whether (...) or (not)’, whereas only 23.21% of if-clauses include similar markers of alternative coordination.

Furthermore, in the case of whether-clauses, the explicit expression of polarity or alternatives is not specifically associated to their use in the [EXT S/O] and the [ADJ COMP] constructions, but is rather an inherent property of the subordinator whether that can be attributed to its historical origin as a comparative form of pronoun who with the specific meaning ‘which of the two’. In fact, whether clauses usually have this alternative value even when they function as adverbial conditionals as in the following examples from our corpus:

(76)  This is true whether you are in a desert or a forest or swimming on a coral reef. <BLX 26>

(77)  Expansion or reformulation should be possible whether or not the first result is positive or negative. <H0S 243>

(78)  Much of the information is relevant whether you are flying a hot-air balloon, a glider or a plane. <BNV 648>

These examples seem to argue against the traditional association of the polar or alternative meanings expressed by whether-clauses and their interrogative nature (Quirk et al. 1985: 1053-1054; Huddleston/Pullum et al. 2002: 975-984). In other words, whether-clauses are or can be polar or alternative even when they are not interrogative but clearly conditional (i.e., adverbial) and no question is explicitly or implicitly being raised, as in (76)-(78). This alternative (rather than interrogative) nature of whether-clauses makes them particularly suitable to appear in contexts where options are being considered, discussed, evaluated, and so on, such as the adjectival contexts that we are discussing in this paper, where the options explicitly or implicitly


suggested by the *whether*-clause are evaluated by the adjective in terms of their plausibility, relevance, effect on the subject, etc.

*If*-clauses, as stated, are in general terms much less frequently associated with the explicit expression of alternatives, but they can be said to be implicitly polar.\(^{11}\) However, their explicit or implicit polarity and their alternative meaning seem to be much more closely associated with their use in the [EXT S/O] and the [ADJ COMP] constructions (examples (79)-(82) below) than in the case of *whether*-clauses, since in their adverbial-conditional use *if*-clauses do not seem to convey any polar or alternative connotations (examples (83)-(84)).

(79) Polly blinked, not quite sure if he was serious. <H7W 939>
(80) I’m still not sure if she means it or not. <A74 861>
(81) It is doubtful if Leonard learned much law from Scott. <A0P 1091>
(82) It is not clear if this refers to his own experiment or to the fact that Jones had by then made his work known to them. <CER 1383>
(83) It’s pretty good fun, but it’d be better if I could float. <A74 394>
(84) I’m especially happy if there are people I love around me. <H94 3009>

Thus, it would appear that in the case of *if*-clauses the polar or alternative meaning is not an inherent property of the clause itself, but is rather construction-dependent and only appears when the *if*-clause is syntactically related to a preceding adjective (or verb, as in *I wonder if*...).

---

\(^{11}\) The implicit polarity of *if*-clauses – and the same applies to *whether*-clauses – can be easily proved by the simple test or adding *or not* at the end of the clause, something impossible with, for instance, *that*-clauses:

i. He was **not sure if/whether** Miss Rossetti would have approved of Ash’s theology [*or not*].
ii. He was **not sure that** Miss Rossetti would have approved of Ash’s theology [*or not*].

It is worth noticing that this contrast in acceptability has nothing to do with the supposed interrogative nature of *if/whether*-clauses, as the two examples above seem to have the same underlying question: ‘Would Miss Rossetty have approved of Ahs’s theology?’.
5. Conclusions

Our corpus-based study of the different patterns where adjectives co-occur with *whether/if*-clauses has revealed a number of significant aspects, most of which deserve a much more detailed analysis than could be offered in this exploratory paper.

In the first place, *whether*-clauses have been shown to be much more frequently found in construction with predicative adjectives, either as their (extraposed) subjects/objects or as their complements within the structure of complex adjective phrases. *If*-clauses, on the other hand, are much more frequently merely adjacent to the adjective and in construction with the whole clause as conditional adverbials. This represents a clear asymmetry in the syntactic behaviour of both kinds of clauses in an adjectival context which diverges from their frequently assumed interchangeability as verbal complements.

The syntactic relation that holds between adjectives and *whether/if*-clauses in the [EXT S/O] and the [ADJ COMP] constructions has been shown to correlate with a series of lexico-grammatical associations: co-selection of the adjective from specific semantic sets, negative or non-assertive context and the explicit or implicit reference to alternatives.

The large majority of adjectives that are more frequently co-selected with *whether/if*-clauses belong to the semantic categories we have informally termed (UN)CERTAINTY (*obvious*, *(un)certainty*, *(un)clear*, *(un)sure*), and OPINION (*arguable debatable, doubtful*). It is this kind of adjectives (and the related verbal expressions (*not*) *know*, *wonder*, *decide*, *doubt*, etc.) that have traditionally prompted the description of this use of *if/whether* clauses as interrogative, assuming that they always imply some kind of underlying question. However, we have shown that, on the one hand, even *that*-clauses may be said to suggest an underlying question in similar contexts and, on the other hand, *whether/if*-clauses also collocate with adjectives whose interrogative value is more than questionable, such as those expressing the (IR)RELEVANCE of some possibility or possibilities (*immaterial*, *(ir)relevant*, *(un)important*), or those expressing concern
over the potential materialization of certain possibilities or alternatives (*bothered, worried, (un)concerned*).

Despite their specific differences, all the adjectives found in construction with *whether/if*-clauses share a specific semantic feature that transcends their own specific lexical content: their negative or non-assertive connotations instantiated by purely lexical, morpho-lexical or, more frequently, lexico-syntactic means. The discussion of the different negative or non-assertive contexts that attract *whether/if*-clauses has proved to be particularly illustrative of the complex interactions between lexical and grammatical features in line with former attempts to show the mutual attraction between specific syntactic classes, specific lexical sets and abstract semantic categories such as negation.

Finally, we have shown that, when they are used in construction with adjectives, both *whether*- and *if*-clauses tend to express alternatives, either explicitly (by means of correlative coordination with ‘… or (not)’) or implicitly. In the case of *whether*-clauses this is an inherent property of the clause itself and might even justify their higher frequency (as compared with *if*-clauses) in adjectival contexts, where, as we have seen, evaluation of explicit or implicit alternatives is an important factor. *If*-clauses, on the contrary, only express alternatives when used in adjectival (and verbal) contexts, but not in their adverbial-conditional uses.

Taken together, the lexico-grammatical associations of the ‘adjective + *whether/if*-clause’ constructions suggest a characterization of *whether/if*-clauses that deviates from their traditional description as (subordinate) interrogative clauses. The data presented in this paper suggest that rather than being concerned with the expression of an underlying or implicit question, *whether/if* clauses are concerned with the (subjective) evaluation of situations or states of affairs presented not as actualized facts but as options or alternatives to choose from – particularly in the case of *whether*-clauses – or as mere unactualized possibilities that are evaluated by means of a predicative adjective in terms of their perceived or assumed viability or relevance, or in terms of their consequences for the subject were they to become real. In other words, in the adjectival contexts discussed in this paper, *whether/if* clauses, rather than interrogative, are essentially modal non-factual
clauses representing in English one specific instantiation of the so-called *irrealis mood* that in other languages is frequently realized by means of verbal modifications, such as subjunctive mood in Spanish or other modal markers in languages like Russian, Finnish or Romanian (Chung/Timberlake 1985). This interpretation can be easily extended to the use of *whether/if*-clauses as verbal complements and, most importantly, it could also accommodate their adverbial-conditional use. Many authors\textsuperscript{12} have noticed the cross-linguistic semantico-pragmatic link between ‘(indirect) questions’ and ‘conditions’, supported at a formal level by the fact that they share some subordinators such as English *if* or Spanish *si*. We believe that the re-interpretation of *whether/if*-clauses as non-factual modal clauses could satisfactorily provide a unified account of all their apparently unrelated uses.

On a more general note, we have shown that the different grammatical constructions associated with the formal pattern ‘adjective + *whether/if*-clause’ are a perfect example of what has been called “lexical grammar” (Sinclair 2000) or the “inseparability of lexis and grammar” (Schulze/Romer 2009: 1). In this respect, we have seen that the most relevant features of the ‘adjective + *whether/if*-clause’ constructions cannot be solely attributed to lexical or grammatical features, but to the joint interaction of both. These constructions even suggest that some notions that are currently used to account for the lexis-grammar interface (e.g., semantic colligation, collostructions) might have to take a step forward in order to account for the interaction of certain syntactic clause types (such as *whether/if*-clauses) and combinations of lexical classes with more abstract categories such as negation or polarity.

References


Davies, Mark 2004-. BYU-BNC. (Based on the British National Corpus from Oxford University Press). <http://corpus.byu.edu/bnc>.


Valera-Hernández, Salvador / Rizo-Rodríguez, Alfonso 1998. A LOB-
Corpus-based Semantic Profile of the Adjective in English Sup-
plementive Clauses. *International Journal of Corpus Linguis-
tics*. 3, 251-278.

Publications.
Exploring Hoey’s notion of *textual colligation* in a corpus of student writing

Abstract

Hoey (2005:13), in his radical new theory of the lexicon, *Lexical Priming*, proposes that ‘[e]very word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations’. He argues that such primings are tied to contexts, and, therefore, it should not be assumed that the priming that operates in one textual domain will operate in another textual domain. I explore Hoey’s notion of *textual colligation*, by examining the textual locations and environments of a variety of single and multiword items in a corpus of undergraduate assignments, derived from the *British Academic Written English (BAWE)* corpus. I identify where the items occur in the text, and I then examine their textual environments, both at a sentence level and at a broader textual level. To test the notion of textual colligation, I discuss the evidence for disciplinary variation and whether words or phrases are primed with different textual colligations, depending on discipline and genre. I argue that textual colligation is bound up with experiences and understandings, not just of how language is used, but also of how texts are organised.

---

1 Author’s email address: p.thompson@bham.ac.uk; affiliation: Centre for Corpus Research, U. of Birmingham. The data in this study come from the *British Academic Written English (BAWE)* corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).
1. Introduction

Michael Hoey’s (2005) theory of lexical priming aims to account for what is natural in language, rather than simply what is possible (contra Chomsky), and one of the core concepts in his theory is that of textual colligation. Fundamentally, the proposition underlying this concept is that “every word is primed to occur in, or avoid, certain positions within the discourse” (Hoey 2005: 13). It is my contention in this paper that Hoey left the concept of textual colligation somewhat underdeveloped and I intend to test and extend the notion through an exploration of student essays taken from a corpus: the British Academic Written English (BAWE) corpus. I will propose firstly that the focus should not be solely on individual words but should extend to pairings of words and other sequences (contiguous and non-contiguous), both lexical and semantic (cf. Hunston 2008), and I will also argue that textual colligation can be understood as part of genre knowledge and of how texts are structured: that is, that writers develop an awareness of the rhetorical patterning of texts and of the language that performs functions within the rhetorical patterns through repeated encounters with texts of a common nature. This awareness may not be explicit; what Hoey’s theory offers us is the perception that language users develop a sense of how words work in combination with each other and in context through their experience of language in use.

Whilst the main purpose of this paper is to test and revise the concept of textual colligation, I intend also to explore the writing of undergraduate students and see whether a corpus investigation can provide evidence of development in student writing across the three years of (UK) undergraduate study. In a previous paper (Thompson 2009), I made an initial investigation of student writing development in Economics and Engineering texts across the three years of undergraduate study which indicated that the writing does become more sophisticated over time. Bereiter/Scardamalia (1987) postulated that writers develop from what they call ‘knowledge-telling’, the reproduction of knowledge, to ‘knowledge transformation’, where they are able to assimilate and critically change the ideas of others into their own
2. Textual Colligation

2.1. Lexical Priming (Hoey 2005)

Michael Hoey’s theory of lexical priming is radical: inverting the conventional perspective of language that places grammar at the centre, he argues that it is lexis that is systematically structured and grammar is the outcome of this lexical structure. Furthermore, he proposes that humans learn through repeated encounters with language in diverse contexts and co-texts. As one observes words typically co-occurring with other words, for example, or in lexico-grammatical patterns, in certain types of context, one’s mental schemata (my phrasing, not Hoey’s) adapt to incorporate this new knowledge. In other words, lexis is cumulatively loaded with the contexts and co-texts in which one has encountered the words. These different pieces of information that one internalises about a word are what Hoey terms ‘primings’ (in a different sense from how the word is used in psychology): in similar contexts the word is likely to occur with particular other words (collocates), particular semantic sets (semantic associations), related to particular pragmatic functions (its pragmatic associations) and with certain grammatical functions (its colligations).

The theory of lexical priming aims to explain psychological phenomena but it is supported by analysis of textual evidence. Although the theory postulates that each person’s understanding of language is to an extent unique as it depends on experience, Hoey argues that corpora can act as evidence of language use on a massive scale which, in turn, allows us to see how systematic are the patterns of collocation and colligation in discourse. He also posits a set of primings at a textual level, among which is that of textual colligation, which he expresses as follows:
Every word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations. (Hoey 2005: 13)

Moreover, he adds:

Such primings are tied to contexts and therefore it should not be assumed that the priming that operates in one textual domain will operate in another textual domain. (Hoey 2005: 13)

A simplistic example of this would be the word *once*, which is primed, among other things, to be part of a fixed phrase *once upon a time*, which in turn is primed to occur in text-initial position for a narrative, typically a fairy tale. Hoey takes a less obvious example to support his argument with the word *sixty*. Looking in his corpus of *Guardian* news text, he observes the following:

- There are 307 instances of *sixty* in his newspaper data.
- Of these, 208 are thematised.
- 200 appear as the first word of the sentence.
- And 14% of the instances are text-initial.

In other words, nearly two thirds of the occurrences of *sixty* are sentence-initial and thematised, which indicates a strong tendency for the word to begin sentences and to be thematised. As with most corpus observations, this indicates probabilities – that words and phrases tend to behave in certain ways.

It is also worth noting that the majority of the instances that Hoey examines are of *sixty* in a temporal clause such as *Sixty years ago*... or *Sixty days after*.... As noted above, Hoey has emphasised that textual colligations may be specific to certain textual domains and, in this case, the textual domain is that of a British broadsheet newspaper, or, to be more precise, the *Guardian* newspaper. There appear to be some constraints also placed on writing practice by the newspaper’s housestyle. The *Guardian* newspaper appears to use *sixty* in headlines (for example, ‘Sixty years ago on Wednesday, the first singles chart
was published in Britain – turning pop music into a competitive sport’

was published in Britain – turning pop music into a competitive sport’)\(^2\) with the option of changing to the digital form within the body of the article (‘And on 14 November 1952, exactly 60 years ago today, NME ran the first singles chart’.). In The Times, in comparison, it seems to be acceptable to use numbers in the title of an article (‘60 years on, the peak of achievement’).\(^3\) The article in The Times contains within its body the sentence ‘By 11.30 this morning 60 years ago, their lives, mountaineering and Nepal itself had changed for ever’, where sixty no longer appears strictly in sentence-initial position but it is still contained within the temporal clause, which is in sentence-initial position. Perhaps a more useful way to describe the tendency is to say that the word sixty (or 60) in newspaper stories often appears within what Hunston has called a ‘semantic sequence’ that functions as a time adverbial adjunct to a sentence, at the beginning of a sentence (the same can be said of seventy or of forty). In these instances, sixty tends to collocate with years and to appear in phrases such as sixty years ago or sixty years on.

A revised statement of Hoey’s proposition, then, might read:

Every word is primed, in its collocations, colligations, semantic associations and pragmatic associations, to occur in, or avoid, certain positions within the text.

This revised proposition has the advantage of accounting for the tendency of sixty to appear in sentence-initial position in a sentence such as ‘Sixty years ago, in the summer of 1950, a small Californian business was preparing to introduce the world to a new musical invention’, but not in ‘The joy of sex at sixty’ (headline from The Independent, 16 May 1999).

\(^2\) See <http://www.guardian.co.uk/music/2012/nov/13/sixty-years-uk-charts#ixzz2UghcOFPe>.

\(^3\) See <http://www.thetimes.co.uk/tto/news/world/asia/article3777393.ece> (subscription needed).
2.2. Textual distribution of phraseological items in student writing
(O’Donnell/Römer 2010)

Hoey’s theory of lexical priming has attracted much attention (111 citations to date, according to Scopus), but textual colligation has only been explored in one study that I am aware of. O’Donnell/Römer (2010) investigated the textual distribution of lexical sequences with a span of 3-5 items, both as fixed sequences (n-grams) and as P-frames (n-grams in which one internal item is variable, such as *at the * of) in a corpus of proficient student writing, the MICUSP (Michigan Corpus of Upper-level Student Papers). The MICUSP corpus contains pieces of writing by students in a variety of disciplines and at different levels (final year undergraduate and graduate students). O’Donnell/Römer’s (2010) study identified n-grams and P-frames that were key to particular disciplines, such as the following:

| the psychological well-being – key to Psychology |
| it has been * that – key to Biology |

Of particular interest to us, they also looked at the position of the n-grams and P-frames across texts, and the method they used was to identify whether a sequence occurred in the first 25% of a sentence, the middle 50% or the final 25%, in order to determine whether the sequence, viewed across all its instances, had a tendency to occur in sentence-initial position, sentence-medial or sentence-final. In addition, they assessed whether the sequence had a tendency to appear in the first sentence of a paragraph (paragraph-initial), the final sentence (paragraph-final) or in sentences that are not first or last (paragraph-medial). A third level of the analysis was to determine whether the phrase tended to appear near the beginning of the text, in the middle or near the end.

---

5  For more information, see <http://micusp.elicorpora.info/teachers>.
The analysis revealed some evidence to indicate that phraseological items are distributed differently, and that some phrases have clear preferences. The item *it would be*, for example, has a strong preference for sentence-initial and text-final positions, while *in order to* *the* has a strong preference for sentence/paragraph-initial positions and avoids sentence-final position.

3. The *British Academic Written English (BAWE)* corpus

The *British Academic Written English* corpus contains 2,761 written assignments from four British universities. The corpus was created as part of a project that investigated the genres of assessed student writing in UK higher education. Data were collected from a broad sampling of disciplines in four disciplinary domains: Arts and Humanities, Life Sciences, Social Sciences, and Physical Sciences. Furthermore, assignments were collected from first-year, second-year, third-year and postgraduate students. The overall sampling frame is shown in Table 1; although the figures are not exactly balanced, the corpus does provide a reasonably balanced sampling (in terms of size) across the domains and the years.

<table>
<thead>
<tr>
<th>LEVELS</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ARTS AND HUMANITIES</strong></td>
<td><strong>Assignments</strong></td>
<td>239</td>
<td>228</td>
<td>160</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td><strong>Words</strong></td>
<td>468,353</td>
<td>583,617</td>
<td>427,942</td>
<td>234,206</td>
</tr>
<tr>
<td><strong>LIFE SCIENCES</strong></td>
<td><strong>Assignments</strong></td>
<td>180</td>
<td>193</td>
<td>113</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td><strong>Words</strong></td>
<td>299,370</td>
<td>408,070</td>
<td>263,668</td>
<td>441,283</td>
</tr>
<tr>
<td><strong>PHYSICAL SCIENCES</strong></td>
<td><strong>Assignments</strong></td>
<td>181</td>
<td>149</td>
<td>156</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td><strong>Words</strong></td>
<td>300,989</td>
<td>314,331</td>
<td>426,431</td>
<td>339,605</td>
</tr>
<tr>
<td><strong>SOCIAL SCIENCES</strong></td>
<td><strong>Assignments</strong></td>
<td>207</td>
<td>197</td>
<td>166</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td><strong>Words</strong></td>
<td>371,473</td>
<td>475,668</td>
<td>447,950</td>
<td>704,039</td>
</tr>
</tbody>
</table>
Our interest in this paper is with whether there is evidence in this corpus of textual colligation. As Hoey emphasises, textual colligation is tied to contexts and it is likely to be tied to textual domains. In this instance, we will focus attention on genre family as a form of textual domain. Nesi/Gardner (2012), in their analysis of the texts in the BAWE corpus, identified 13 genre families, one of which is the essay. This is the genre family of assessed writing that is most frequently found in the arts and humanities and in social sciences (although it is worth remarking that it is not the only one). The social purpose of the essay, according to Nesi/Gardner (2012), is to develop the ability to construct a coherent argument and develop critical thinking skills. Members of this genre family typically feature an introduction, a series of arguments and a conclusion. Within the family of ‘Essay’, they discern six essay genres:

1. discussion (issue, pros/cons, final position);
2. exposition (thesis, evidence, restate thesis);
3. factorial (outcome, conditioning factors);
4. challenge (opposition to existing theory);
5. comparison (series of comparative points or arguments);
6. commentary (series of comments on a text).

There are 1,237 essays in the BAWE corpus, making up nearly 50% of the corpus contents, which provides us with a reasonably large quantity of data for analysis, and these can further be distinguished as first-year, second-year and third-year essays, for purposes of investigation of change across the years.

---

6 Source: <http://tinyurl.com/pylo54g>.
The corpus files are marked up for a range of formatting and structural features, and here we are interested in the markup which identifies the number of sentences within a given paragraph, and also the number of paragraph out of however many paragraphs there are in that given text (details are given in the following section). This information can act as a form of textual coordinate: the information helps one to see the location of words or phrases relative to the number of sentences in a paragraph and the number of paragraph in the text. These textual coordinates can potentially help us to identify whether or not words and phrases have a tendency to appear in certain parts of texts; in this case, we will see if they tend to appear in sentence-initial position, and in paragraph-initial sentences or in paragraph-final sentences. Using the textual coordinates information, we will address the following question:

Is there evidence of textual colligation in essays in the BAWE corpus as a whole, and in subsets according to level?

We will do this by looking at individual lexical items, identified by drawing up lists of keywords, and their collocates and colligates, and then we will also see how certain contiguous sequences (n-grams) and non-contiguous sequences (P-frames, following O’Donnell/Romer 2010) are used in student essays.

4. Methodology

Paragraphs can be of varying lengths: from the shortest paragraph, consisting of one sentence only, to the longest, which can contain 30 or more sentences. The one regularity for paragraphs that are two or more sentences in length is that there will be an initial sentence and a final sentence in the paragraph. Therefore, if we are interested in comparing paragraph position across widely varying set of paragraph links, it makes sense to focus attention on the paragraph-initial and paragraph-final sentences only and to treat all others as paragraph-
medial. If a paragraph has four sentences, it will have an opening and a closing sentence, and if a paragraph has 35 sentences, it too will have an opening and a closing sentence.

The BAWE corpus is available as a collection of text files or XML files; for this investigation the XML files were used. The XML files are marked up for presentational features following the TEI Guidelines and the relevant part of the markup is the indication and numbering of paragraphs of sentences within paragraphs. Each paragraph is identified with a `<p>` element and with an ‘n’ attribute that indicates that a given paragraph is, for example, the fifth paragraph out of 26 paragraphs in a text, and the first sentence in the paragraph will then also be given an ‘n’ attribute which identifies it as number 1 out of, say, three sentences. To illustrate this, here is an example taken from the BAWE corpus:

(1) <p n="p5.26">
<s n="s1.3;p5.26"> Demographic changes illustrate that marriage is increasingly becoming seen as outdated and is being supplanted by cohabitation. </s>
...

The headers for the files contain metadata for the file, amongst which is a classification of the text’s genre. For this study, we needed to work with the files in a variety of programmes and so the first steps were to identify and extract all files from the larger corpus that were essays, and then to create new subcorpora which were sentences from the essay files that were paragraph-initial, paragraph-medial or paragraph-final. To this end, all sentences which were numbered ‘1’ in paragraphs containing more than one sentence were treated as paragraph-initial, those which were sentence x out of x sentences (where x>1) were treated as paragraph-final, and the rest (also excluding 1.1 – one out of one sentence in a paragraph) were treated as medial. The process was semi-automatised (using regular expression searches), but care had to be taken because it soon became obvious that some of the sentence-boundary scripts used in the annotation of the BAWE corpus were not perfect: in the Law essays, for example, some sentence boundaries were created after the full stop in Vs. (e.g., Bell Vs. Mayweather), since the scripts had relied on punctuation marks to decide
on the ends and beginnings of sentences. Other problems were caused by poor punctuation in the original texts (lack of full stops, insertion of full stops in mid-sentence, etc). For simplicity’s sake, these were deleted from the corpora.

The next stage was to identify the key words for the paragraph-initial and paragraph-final sentences. Wordlists were created in AntConc 3.2.4w (Anthony 2011) and then the paragraph-initial and paragraph-final wordlists were compared against the composite file containing all sentences using the Keywords tool. Typically, the Keywords tool is used to compare a small corpus against a large (often generalised) reference corpus, but in this case the purpose of the analysis was to identify words which occur more frequently within the subcorpus than would be expected, given the frequency of occurrence within the corpus as a whole.

The final stage involved identifying the most frequent P-frames in the complete essay dataset, and this was done using William Fletcher’s freeware program, KfNgram. The program identifies n-grams and then, based on that information identifies sequences from the n-gram list in which n-1 out of the items in the sequence remain constant while a single item is variable: these are P frame candidates. Having identified the most frequent P-frames, we then calculated their relative proportional normed frequency so that we could see whether any of the frames tended to occur more frequently in paragraph-initial or-final sentences.

5. Analysis

5.1. General

As can be seen in Table 2, there are approximately 15,000 each of paragraph-initial and of paragraph-final sentences, and the number of paragraph-medial sentences is roughly 3.5 times greater than either the number of paragraph-initial or paragraph-final sentences. In the results shown below, the frequencies shown are normed frequencies which take into account the relative size of each data sample.
Table 2 shows the number of sentences in the three subcorpora (paragraph-initial, paragraph-medial and paragraph-final). The right-most column shows the ratio: 1:3.5:1. The numbers of paragraph-initial and paragraph-final sentences do not match up because many instances were deleted due to faulty sentence identification.

The key words for paragraph-initial and paragraph-final sentences are shown in Table 3. It should be remembered that these figures were derived by comparing one specific set of sentences in one category (for example, paragraph-initial) to the whole corpus (comprising paragraph-initial, paragraph-medial and paragraph-final sentences). Some of these words are topic-specific (*century, modality*), while others describe either the genre itself (*essay*) or elements and relations in argumentation (*conclusion, problem, argument; thus, therefore, hence*). The elements in argumentation occur in paragraph-initial sentences, while markers of relation occur in paragraph-final sentences. There is also one word which is key in both paragraph-initial and paragraph-final sentences (*essay*).

Another point to be made about paragraph-final sentences is that there are five modal auxiliaries amongst the top 20 keywords: *will, shall, may, should, would*.
Table 3. Keywords in paragraph-initial and paragraph-final sentences in the *BAWE* essay data.

Table 3 shows the keywords in paragraph-initial and paragraph-final sentences in the *BAWE* essay data. The ‘hits’ column shows the raw frequency of occurrence, and the ‘keyness’ column shows the AntConc Keywords scores (based on log-likelihood calculations).

In order to explore the possibility that words are primed to occur in certain parts of the discourse (in this case in paragraph-initial or paragraph-final sentences), we will look in the following sections at two keywords from the paragraph-initial column and two from the paragraph-final column: *essay*, *important*, *thus*, and *therefore*. We will look at the modal verbs in a later section.

### 5.2. Essay

On the basis of individual word frequency alone, there is no evidence that the word *essay* in any way compares with Hoey’s example of *sixty* (see §2.1 above), which had a clear tendency to occur in theme position within sentences. The word *essay* occurs 298 times in paragraph-initial sentences, 370 times in paragraph-medial, and 265 times in paragraph-final sentences; in other words, there are
tendencies for the word to appear at the beginnings and ends of paragraphs but it does also occur quite frequently in the middle of paragraphs. However, by looking at the words that occur immediately on either side of the word in question, we find that in all the texts, regardless of the position of the sentence within the paragraph, *essay* is immediately preceded by the word *this* 969 times, and in 329 instances the word *this* is capitalised, which indicates that it appears in sentence-initial position. Extending the span further, we find that the proposition *in* occurs 317 times to the left of *this essay*, of which 132 are capitalised (sentence-initial position). This provides some evidence that *this essay* and *in this essay* have a tendency to occur in sentence-initial position (329 + 132 = 461).

Moving now to make some observations about *essay* in paragraph-initial sentences, we find that the word *essay* occurs 419 times, and it is preceded by lowercase *this* 194 times and by upper case *this* 129 times. In other words, more than three quarters (323/419) of the occurrences of *essay* in paragraph-initial sentences are in combination with *this*. *this essay* is preceded by *in* (lowercase) 55 times and *In* (uppercase) 65 times. Summing this up, we can say that in paragraph-initial sentences the combination *this essay* occurs 323 times, of which 120 occur in the longer sequence *in this essay*. *This essay* or *In this essay* appear in sentence-initial position 195 times. There is, therefore, some evidence to suggest that the word *essay* tends to occur in the sequences *this essay* and *in this essay* in paragraph-initial sentences, and that the sequences tend to occur in sentence-initial position.

Another sequence is *of this essay* which occurs 34 times in all, with 26 occurrences of these in paragraph-initial sentences, such as *The purpose of this essay* or *The argument of this essay*. Sixteen out of the 26 instances are in subject position within the sentence.

So far, we have found some support for the notion of textual colligation inasmuch as that we have observed tendencies in the data. We can also infer from the preliminary evidence that the combination of *this* and *essay* occurs in metadiscursive statements about the purpose, the aims or the scope of the essay, and thus that there is a link between the phrasing and the rhetorical function of the sentence.
5.3. Important

We now turn to important, which is the fourth ranked keyword in paragraph-initial sentences, occurring 544 times in this category, and 1,884 times in all sentences. Looking at the 544 occurrences in the paragraph-initial category, we can make the following observations:

- In 141 instances, important precedes a noun (such as factor, feature, aspect, role, issue, point).
- In 144 cases, important is premodified by an adverb (such as most, very, extremely, highly).
- Important is followed by infinitive to 104 times, such as in ‘It + BE + important + to’.
- The pattern important * of is more characteristic of paragraph-initial sentences, with 14 instances of ‘(An|One|Another) important NOUN of’ out of 18 belonging to the paragraph-initial category.

In many cases, the first sentence of a paragraph is a statement of a proposition. In writing textbooks, this is commonly referred to as the topic sentence, particularly in the US. Recently, writing researchers such as Hyland (2003) have argued that the concept of the paragraph-initial topic sentence is grossly overstated, but in the BAWE corpus it is noticeable that paragraphs often do begin with a general proposition which is then elaborated, exemplified and supported in the following sentences of the paragraph, as the following example:

(2) A vitally important difference when considering delinquent acts committed by girls or boys is that for girls’ delinquency becomes embodied, for boys it does not (Smart 1995: 48; Cox 2003: 164)

They also often can contain a reference back to the previous discourse, as in:

(3) It is also important to mention briefly, how the Africans were viewed in North America and how they were viewed in South America, as this reinforces the idea of racism in England and colonial America.
Again, we can see some evidence here of a relationship between the word, its patterning, the functions of those patterns and the word’s distribution in text that relate to rhetorical functions and preferred rhetorical organisation (the positioning of a topic sentence at the beginning of a paragraph).

5.4. Thus and therefore

These two words are key to the paragraph-final sentences. Some initial observations are that:

1. *Thus* occurs 603 times and of these 354 are in sentence-initial position (capital T in *Thus*), approximately 59% (cf. 32% paragraph-medial, 22% paragraph-initial).
2. *Thus* is preceded by *and* (*and thus*) in 126 sentences (approximately 21%; cf. 27% paragraph-medial, 24% paragraph-initial).
3. *Therefore* occurs 880 times and 341 of these are in sentence-initial position, approximately 40% (cf. 32% paragraph-medial, 22% paragraph-initial).
4. *Therefore* is preceded by *and* (*and thus*) in 128 sentences (approximately 20%; cf. 23% paragraph-medial, 26% paragraph-initial).

As these two words serve to show the relationship between statements, they do not form strong associations with words other than with *and*. Of the four observations, there is little that is exceptional about the placement of *and thus* and *and therefore* in the three sets of data, but there is good evidence to support the proposition that in paragraph-final sentences, there is a stronger preference for *Thus* (59%) and *Therefore* (40%) to appear in sentence-initial position (compared to the tendency in the other two groups).

The tendency can be explained, to some extent, by the rhetorical need for writers to conclude the line of argument with a statement of the logical consequence of the propositions in the preceding sentences, and the placement of *thus* or *therefore* in sentence-initial position can
flag up, and emphasise, the logic that results in the final statement. For example:

(4) Therefore it is reasonable to say that it is the reaction of the government that is the deciding factor between crisis and famine, and so can be classified as the ‘cause’.

Other writers, however, choose to place the flag after the auxiliary or the main verb in the main verb phrase:

(5) Business managers must therefore adopt a structured approach to planning business strategies, so that long-term strategic objectives and the events that enhance these objectives can be accomplished simultaneously.

(6) Liberalism thus appears more moral and less pessimistic than Realism, possibly accounting for its popularity in political discourse today.

In all cases the words act as intersentential connectors, showing the relation between the previous sentences and the following, concluding proposition.

5.5. P-frames

We turn now to the consideration of non-contiguous sequences, or P-frames, following O’Donnell/Römer (2011). The first step is to identify the most frequent four-item phrase gram heads (that is, four grams in which one item is variable), and then choose P-frames that are less equally distributed. The results are shown in Table 4.

<table>
<thead>
<tr>
<th>4 PHRASE GRAM HEADS</th>
<th>INITIAL</th>
<th>MEDIAL</th>
<th>FINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>the * of the</td>
<td>31.14</td>
<td>38.74</td>
<td>30.12</td>
</tr>
<tr>
<td>in the * of</td>
<td>32.69</td>
<td>37.72</td>
<td>29.59</td>
</tr>
<tr>
<td>to the * of</td>
<td>35.41</td>
<td>37.31</td>
<td>27.27</td>
</tr>
<tr>
<td>it is * to</td>
<td>44.09</td>
<td>26.24</td>
<td>29.67</td>
</tr>
<tr>
<td>the * of a</td>
<td>26.68</td>
<td>40.86</td>
<td>32.46</td>
</tr>
<tr>
<td>in order to *</td>
<td>31.49</td>
<td>36.21</td>
<td>32.31</td>
</tr>
<tr>
<td>it is * that</td>
<td>31.12</td>
<td>34.21</td>
<td>34.67</td>
</tr>
<tr>
<td>as a * of</td>
<td>31.23</td>
<td>31.73</td>
<td>37.04</td>
</tr>
<tr>
<td>of the * of</td>
<td>29.06</td>
<td>41.71</td>
<td>29.23</td>
</tr>
<tr>
<td>at the * of</td>
<td>37.48</td>
<td>36.24</td>
<td>26.28</td>
</tr>
</tbody>
</table>
Table 4. The most frequent 4 item phrase gram heads in the BAWE essay corpus.

Table 4 contains the most frequent 4 item phrase gram heads in the BAWE essay corpus, showing normed frequency proportions of occurrence in the three subcorpora: paragraph-initial sentences, paragraph-medial and paragraph-final. For each P-frame, the three figures add up to 100 and thus the values show the percentage of occurrences that each subcorpus accounts for. Boldface is used for items which account for more than 50% of instances in one column, or over 40% in two columns.

One of the * occurs with far greater frequency in paragraph-initial sentences, and with markedly less frequency in paragraph-final position. Looking at the data, the following observations can be made:

- Of 763 instances of one of the *, 221 are in sentence-initial position (capitalised ‘one’).
- The words that immediately follow one of the most frequently are: *most (182), main (74), major (32), first (27) and key (25).
There is evidence again that the P-frame *One of the* can be used to begin a paragraph-initial sentence, which may function as a topic sentence. This sentence also can feature a value claim that asserts centrality or priority, and this is closely connected to the development of an argument and to the expression of stance (‘I will explain to you what is key in this debate’). For example:

(7) One of the principal factors in explaining the failure of the mission was the fact that the Aborigines, and indeed the surrounding white population, used the missions for purposes other than religion.

Another use of *one of the* is to enumerate or itemise, as in:

(8) One of the ways in which Orwell articulates the relationship between language and power is through psychological manipulation.

In both examples, the P-frame can be associated with a rhetorical function (asserting centrality and enumerating). The same can be observed with *It is important* where the following word is usually either *to* or *that*. The P-frame is often preceded by an adverbial such as *Firstly* or *However*, and in the latter case, it is possible to see that there is a contrast drawn with the preceding paragraph.

We can move from the P-frame *it is important* to two lexico-grammatical patterns, *it + BE + ADJ + that* and *it + BE + ADJ + to* that have been the focus of studies of stance in academic writing, such as Charles (2000) and Groom (2005), both of which demonstrate how writers in different disciplines use these lexico-grammatical patterns to construct academic *personae*, appropriate to a given discipline. Where *it is important* tends to appear in paragraph-initial sentences, we find that claims of clarity are often associated with conclusions, as in:

(9) In conclusion it is clear that the sources believed that Spartan women differed to their Greek counterparts, but it is impossible to judge whether this was the reality.

(10) In this sense then it is clear that the Germans, at least, felt that the American movement was an important influence on their own program of eugenic racial hygiene.
The first of these examples appears in paragraph-initial position, but is marked as a conclusion by the phrase *In conclusion* (which is key to paragraph-initial sentences) and the second is from a paragraph-final sentence, summing up the point of the paragraph.

In conclusion, this brief exploration of P-frames in the corpus has identified two sequences that play a role in the argumentation of an essay by highlighting claims of importance and also helping the writer to construct an appropriate authoritative, critical *persona*. We will now extend our investigation to a consideration of how these patterns are used across the levels.

5.6. Variation by level

In the figure below, we can see the frequency of use at three levels (first year, second year, third year) of four sequences: *it is * to* (where the asterisk covers a range of adjectives that can fill that space; *be able to * (see Table 4 above); *this essay* (see §5.2 above); and *one of the *.

![Charts showing the frequency of occurrence of four sequences in the subcorpora.](image)
Figure 1 shows four different charts with the frequency of occurrence of four sequences in the subcorpora. The columns in each case are grouped into paragraph-initial (P-initial), paragraph-medial (P-medial) and paragraph-final (P-final), and the columns for each category show first (leftmost) the normed frequency for first-year essays, then the second-year essays, and lastly the third-year essays. If the left column is higher than the others, this indicates a decrease in use, and if the right column is the highest, this indicates an increase.

We can see that the patterns which we identified as characteristic of particular places in paragraphs (this essay, one of the*) are used more frequently by third-year students than first- and second-year ones. In the case of this essay, this is regardless of the place within the paragraph – the frequency increases in initial, medial and final sentences – but in the case of one of the * the movement is towards a greater use of the pattern within paragraph-initial sentences. The evaluative pattern it is * to presents a less clear profile, with the use remaining steady in paragraph-final environments, but increasing steadily in paragraph-initial sentences from first to third year. One can speculate that this is due to an increasing understanding of the need to foreground clear statements of stance in the openings to paragraphs as writers become more experienced.
We remarked above that five of the modal verbs are amongst the key-words for paragraph-final sentences. In Figure 2, we can see how use of these five modal verbs (regardless of paragraph position) changes over the three years. While will and shall drop in the third year (often used to expressed futurity), would, may and should increase from year to year. A likely explanation for this is that student writers gradually learn the importance of qualifying claims, of recognising that diverse views are possible and of hypothesising as they progress through their degrees.

6. Conclusions

In conclusion, we have found evidence to suggest that some words and some phrases tend to occur in paragraph-initial or -final sentences, thus weakly supporting the concept of textual colligation. It has been argued that, in talking of textual colligation, it is necessary to consider the relationship between words and their collocates, their pragmatic associations and their semantic associations. Hoey (2005) says that primings are tied to contexts and we can go further and propose that these may be rhetorical contexts. Our observation that this essay and one of the tend to appear in paragraph-initial sentence position can be related to the rhetorical function of such sentences in indicating the nature of what is being talked about (these are metadiscursive statements). To some extent, this brings us back to a well-known observation, that forms relate to functions (such as ‘focussing’, ‘evaluating’, ‘defining’, etc), but it also suggests that writers can develop a richer repertoire for writing, for example, essays, through experience of such texts, either as a writer or as a reader. And yet it may not be simply through reading of essays that a writer learns to write essays, but also through the reading of comparable genres – after all, how many essays do undergraduate students actually read, other than their own?

There is also some evidence to suggest that the tendency increases ‘year’ by ‘year’ as writers become more experienced in wri-
ting essays, and it can be argued that this occurs as academic writers move from knowledge-telling to knowledge-transformation (cf. Bereiter/Scardamalia 1987). The evidence to support this is that certain patterns associated with evaluation occur more frequently in the third year data than in the other two years, and also that characteristic patterns such as *this essay* and *one of the* increase year by year.

Finally, it is necessary to admit the limitations of the data and the methodology that have been used in the present study. One problem is that the assumption that paragraph-initial and paragraph-final sentences are to some degree units in themselves is questionable. Conclusions to paragraphs, for example, can easily span two or more sentences, rather than just one, and similarly a writer can put forward a topic thesis statement in the first two sentences of a paragraph. The method used in this paper, however, while it may not be precise, does provide a good indication of tendencies. What is more questionable, though, is the extent to which all the texts described as ‘essays’ are comparable, given both that Nesi/Gardner (2012) have described this as a genre family, containing at least six genres, and also that the notion of what constitutes a genre can vary from discipline to discipline. Further research is needed that explores textual colligation in a more defined set of texts, such as in history essays of a set of types, at different levels.

References


Thompson, Paul / Heuboeck, Alois 2009. *Xtractor*. Available upon request from Paul Thompson, p.thompson@bham.ac.uk.
ALEJANDRO ALCARAZ-SINTES has been senior lecturer (tenured) at the University of Jaén since 2011. He has also lectured at various universities (Cergy-Pontoise, Jena, Wurzburg, Tartu, Opole and Toronto) as a guest lecturer. His main research interests lie within English historical linguistics, particularly the complementation of adjectives and medieval medical texts. He has written a monograph, book chapters, and a number of journal articles and reviews, some in international journals, and has co-edited an international monograph. He has co-edited two issues of *The Grove* (2000 and 2002) and the proceedings of the SEDERI (Spanish Society for English Renaissance Studies) and SELIM conferences for 2004 and 2005, respectively, and is guest co-editor of the first issue of *RiCL – Research in Corpus Linguistics*.

MIGUEL-ÁNGEL BENÍTEZ-CASTRO has recently completed his PhD at the University of Granada (Spain). From 2009-2013, he held a four-year PhD scholarship funded by the Spanish Ministry of Education, during which time he developed his research on a multifaceted approach to the analysis of English shell nouns. His research output for the past four years has spanned various areas of linguistic enquiry, ranging from derivational morphology, EAP, learner language and systemic-functional linguistics, all of which he has managed to combine in his study on the syntactico-semantic and textual contribution of shell nouns to different types of discourse. As a research visitor at the Centre for Advanced Research in English at the University of Birmingham (UK) in 2011 and 2012, he acquired expertise in corpus-assisted discourse analysis, where his main research interest lies.

SERHIY BYKH holds an M.A. degree in Linguistics and Computer Science and currently pursues PhD research in Computational Linguistics at the University of Tübingen. His main research focus lies on
linguistic modelling and machine learning in the context of Native Language Identification (NLI), the task of automatically identifying the native language of the writer based on non-native texts. His interests include data- as well as theory-driven approaches to automatic analysis of natural language data with a focus on linguistic feature engineering, both in NLI and more general in the context of related tasks potentially benefiting from combining linguistic insights and the automatic computational analysis of corpora.

**Nuria Calvo-Cortés** is an Assistant Lecturer of English Linguistics in the Department of English Philology I at Complutense University, Madrid (Spain). She received her doctorate in English Philology from Complutense University in February 2012. Her recent publications include “English and Spanish 'Love' Collocations: A Historical Evolution”, *Translation Journal* (2010) and a review of “Astronomy ‘playne and simple’, the writing of science between 1700 and 1900”, *Atlantis* (June, 2013). Her research focuses on the semantic and syntactic evolution of expressions in the history of the English language, with special interest in the Late Modern English period. Currently she is carrying out a study of nominal progressive structures in British English in the 18th and 19th centuries and she is also supervising a doctoral thesis on the development of maritime expressions in English, Spanish and Arabic.

**Jan Čermák** is Professor of English at Charles University in Prague. His research is focused on the history of English (with particular interests in Old and Early Middle English morphology, word-formation and literary language), on Old and Middle English literature (with particular interests in heroic poetry, romance and the Alliterative Revival), and on the history of Finnish (with particular interests in *The Kalevala*). His publications include the first Czech translation and critical edition of *Beowulf* (2003) and the first Czech critical edition of *The Kalevala* (2014).

**Eduardo Coto-Villalibre** holds an MA in English Language and Literature from the University of Santiago de Compostela, Spain and is currently preparing a PhD in English Linguistics at the Department
of English and German Philology at Santiago de Compostela, on a four-year FPU research grant from the Spanish Ministry of Education. He is a member of the Research Unit for Variation, Linguistic Change and Grammaticalization (University of Santiago de Compostela <http://www.usc-vlcg.es/ECV.htm>). His main research interests include corpus linguistics and World Englishes, in particular East and Southeast Asian varieties of English.

THOMAS EGAN is Professor of English Linguistics at Hedmark University College in Hamar, Norway. His research interests encompass topics within the areas of corpus linguistics, contrastive linguistics, cognitive linguistics and historical linguistics, including grammaticalisation. He has made numerous conference presentations and published widely within these fields. In recent years he has concentrated for the most part on contrastive studies of constructions containing propositional phrases in English, French, Norwegian and Swedish. He has also written a monograph on complementation, entitled Non-finite complementation: a usage-based study of infinitive and -ing clauses in English, published by Rodopi in 2008. Egan’s postal address is Hedmark University College, LUNA, Post box 4010 Bedriftssenteret, 2306 Hamar, Norway.

TERESA FANEGO has been Full Professor of English Language and Linguistics at the University of Santiago de Compostela (Spain) since 1990 and is the director of the Research Unit for Variation, Linguistic Change and Grammaticalization. She has published widely on the language of Shakespeare and his contemporaries, and on various aspects of English historical syntax since Middle English times, including the system of sentential complementation and the history of English verbs of motion. Between 2005 and 2013 she served as Editor-in-Chief of the scholarly journal Folia Linguistica, published by Mouton de Gruyter for the Societas Linguistica Europaea.

JULIA KRIVANEK holds M.A. degrees in Computational Linguistics and in Theoretical and Applied Linguistics. She has worked as a researcher at the University of Tübingen in several projects, including the CoMiC project (SFB 833/A4) investigating the automatic content
assessment of answers to reading comprehension questions and the EU LLP project MERLIN providing a Multilingual Platform for the European Reference Levels. Currently she is working as a computational linguist at one of the leading providers for technical translation in Germany. Her research interests include corpus linguistics, automatic text analysis using a wide range of linguistic features, automatic text classification, as well as computer-assisted and machine translation and quality assurance analysis for translation.

MARÍA JOSÉ LÓPEZ-COUSO is Senior Lecturer at the University of Santiago de Compostela, Spain, and a member of the Research Unit Variation, Linguistic Change and Grammaticalization. Her main research areas are morpho-syntactic and pragmatic change and grammaticalization processes in the history of English, with specific interests in the fields of clausal complementation, existential constructions, negation, and parentheticals. She has published extensively on these and other topics in various collective volumes by international publishers and journals like English Language and Linguistics, Folia Linguistica, Journal of Historical Pragmatics, Folia Linguistica Historica, and Journal of English Linguistics. She has also co-edited several volumes in the areas of historical linguistics (English historical syntax and morphology, John Benjamins 2002; with Teresa Fanego and Javier Pérez-Guerra; Information structure and syntactic change in the history of English, OUP 2012; with Anneli Meurman-Solin and Bettelou Los), and grammaticalization studies (Rethinking grammaticalization: New perspectives; John Benjamins 2008; with Elena Seoane); Theoretical and empirical issues in grammaticalization, Benjamins 2008; with Elena Seoane).

BEATRIZ MATO-MÍGUEZ holds an MA in English Language and Literature from the University of Santiago de Compostela (Spain) and is currently preparing a PhD in English Linguistics at the Department of English and German at Santiago de Compostela, on a four-year FPU research grant from the Spanish Ministry of Education. Her main research interests include the interconnections between grammar and discourse, with a special focus on insubordinated conditional clauses and their differences in usage in British and American English.
BELÉN MÉNDEZ-NAYA is Senior Lecturer at the University of Santiago de Compostela, Spain, and a member of the Research Unit Variation, Linguistic Change and Grammaticalization. Her main research interests are grammaticalization processes in the history of English, clausal complementation, the development of English intensifiers, parentheticals and the emergence of pragmatic markers. She has contributed to a number of collective volumes by publishers like Mouton and John Benjamins, and has published her work in peer-reviewed international journals such as English Language and Linguistics, English Studies, Folia Linguistica Historica, Journal of English Linguistics, Journal of Historical Pragmatics, and Neuphilologische Mitteilungen. She has edited the volume Sounds, words, texts and change (John Benjamins 2002; with Teresa Fanego and Elena Seoane) and a special issue on English intensifiers (English Language and Linguistics 12.2, 2008).

DETMAR MEURERS is a Professor of Computational Linguistics at the University of Tübingen, which he joined in 2008 after eight years as a faculty member at The Ohio State University. His current research involves a range of topics interfacing with linguistic modeling, such as issues in intelligent language tutoring systems (TAGARELA), visual input enhancement of learner-selected web pages (WERTi), information retrieval for language learners, automatic content assessment for reading comprehension questions (SFB833-A4), task-based learner corpora, annotation schemes for learner corpora, corpus annotation error detection (DECCA), the use of corpora for theoretical linguistics, and the interface of syntax and information structure.

MATTI RISSANEN is Professor Emeritus of English Philology at the University of Helsinki. His main field of interest is corpus-based analysis of the history of English syntax and lexis, from Old to Late Modern English. He was the leader of the project compiling The Helsinki Corpus of English Texts (publ. in 1991) and he wrote the chapter on syntax in Volume 3 of The Cambridge History of the English Language (1999). In recent years his research has been focused on the development of adverbial connectives and he has published a number of articles on this topic, paying particular attention to grammaticalisation
and language contact. He participated as a consultant in the Shakespeare Translation Project in 2004-2013. He is an active member of the Research Unit for Variation, Contacts and Change in English (VARIENG), at the University of Helsinki.

JUAN SANTANA-LARIO is Senior Lecturer in the Department of English and German at the University of Granada, Spain, where he has taught a variety of courses on English language and linguistics since 1987. He has also taught at a number of universities including, among others, Williams College (Massachusetts) and the University of Delaware. His current main areas of research include English morphology and syntax, corpus linguistics and forensic linguistics. He has over 35 years experience in corpus research and currently works on its application for research on morphological and syntactic patterns and also applications in the classroom. He is one of the organizers of Talc6 (2004, Granada). He is the author of *La categoría gramatical de pronombre en inglés* and co-editor of *Corpora in the Foreign Language Classroom*.

PAUL THOMPSON is the Director of the Centre for Corpus Research at the University of Birmingham. He took up this post in 2009, after working as a lecturer for thirteen years at the University of Reading, and for fourteen years in Japan. His research interests are in academic and other specialised discourses, in the linguistic aspects of human-computer interaction, in uses of educational technologies in language learning, and in the exploitation of corpus resources and methodologies in learning about language. He has worked on four nationally-funded corpus research projects, the most recent of which is the ESRC-funded project 'Investigating Interdisciplinary Research Discourse: the case of Global Environmental Change', on which he is the Principal Investigator. This project is a collaboration with the scientific publisher Elsevier. He is a former Secretary of the British Association for Applied Linguistics (BAAL), and has been Co-Editor of the *Journal of English for Academic Purposes* since 2009. He has published many papers and edited several collections on corpus and applied linguistics. Most recently he co-edited, with Ana Díaz-Negrillo

Ondřej Tichý is a PhD candidate in English Linguistics at Charles University in Prague (supervised by Jan Čermák). His main research interest is in the history of English and the application of corpus and computational linguistics in the field of language diachrony. He is the administrator of the online *Anglo-Saxon Dictionary* by Bosworth & Toller (bosworthtoller.com) and he has recently completed the project of the Automatic Morphological Analyser of Old English.

Salvador Valera-Hernández has been senior lecturer at the University of Jaén (1998-2012) and at the University of Granada (2012 to date), and teaches regularly at Pavol Jozef Šafárik University in Košice, Slovakia (2006 to date). He has supervised a number of publicly funded PhD projects, has managed a number of publicly funded research actions, and has organized several international events in Spain and abroad since 2002. He has also undertaken publicly funded research leave in frontline Departments of several universities (Victoria University of Wellington in New Zealand, Leeds and Vienna) under the supervision of renowned scholars such as Prof. Laurie Bauer and Prof. Dieter Kastovsky. His main field of interest is the morphology of English and has published books and papers on these issues with some of the major publishers, both as a (co-)editor and as a (co-)author. He has been reviews editor of Edinburgh University Press’ journal *Word Structure* since 2008.
Subject index

abstract 39, 70, 288, 292-293, 297
~ion 174, 285, 293-296, 298, 308-309
see also under category, meaning, noun
Acquisition 67, 165, 287-288, 290, 293
Second Language 285-286, 299
active 207-210, 213-214, 222
~isation 211
attributive 211
comparative 146, 154
participial 209-211, 213, 223, 228
predicative 92, 94, 168, 325-328, 340-341
superlative 154
see also under passive, phrase
adjunct 90, 185, 208, 351
conditional 261
clause 87, 104, 263, 317, 322
conditional 317, 321-322, 338-342
degree 211
see also under connective, phrase, subordination
adversative 205, 212, 214, 217-218, 224, 227-228
affix 60, 164
see also prefix, suffix
see also under passive, phrase
agent 207-211, 214-215, 221-222
(in)animate 63, 207-208, 214, 222
(non-)~ivity 63, 208, 212-213, 215-216
by~ 167, 205, 208-211, 214-215, 221-222, 228
gradeence 216
(non-)human 207-209, 222
Janus 208, 210
~less 208, 214, 221
~like 208, 210
(non-)specific 209
see also under passive, phrase
agglutinating 156
as 300-301
binary 301, 303
dative 299, 307-308
locative preposition drop 306-307
multi-class (subcat) 303-304
simple reciprocal 300
single-class 303-304
syntactic 165, 293, 299, 301-302, 309
ambiguity 29, 38, 40-41, 45, 48-50, 74, 88, 90-95, 146-147, 225, 274
analogy 67, 74
analyticity 18, 32, 142, 153, 155-156
grammatical 141-142
index 142
anaphora 167, 171, 174-177, 185-186, 191-192, 197
(in)animate 120, 167, 205, 207-208, 213-214, 216-218, 222, 225-228, 301
see also under agent
automatic 164
error 285
manual 285, 319
multi-level(/-layer) 165, 292, 308
parts-of-speech 114, 289, 291
pragmatic 288
semantic 288
syntactic 114, 288, 299-301
antecedent 173, 176, 184-186, 191-193, 197
anticipator(y) 100, 102, 326-328
apodosis 87, 260-261, 264
argument 302, 307
semantic 70-71
syntactic 70, 300
article
(in-)definite 177, 185, 188-189, 196
zero 188-189, 196
(non-)assertiveness 88, 98-99, 104, 168, 319, 331, 334-335, 341
see also under context, matrix
association 60, 121, 192, 196, 207, 222, 315, 321, 325, 338-339, 362
lexico-grammatical 315, 330-331, 340-342
pragmatic 349, 351, 368
semantic 349, 351, 368
attrition 32
beneficial 205, 213, 217-218, 224, 228
‘betweenness’ 236, 242, 253
borrowing 16, 62, 66, 117, 119, 125-126, 132
boundary 92, 101, 104, 189, 246, 295, 297
~crossing 185, 245-246
see also under sentence
Bulgarian 294
can 265, 270
cataphora 167, 171, 176-177, 185, 191-192, 197
abstract 165, 315, 337, 341-342
~ization 61, 68, 145, 174, 239, 265, 292
formal 148
functional 146
grammatical 288, 331
lexical 288, 301, 331
paradigmatic 144-146, 149
part-of-speech 165
semantic 168, 223, 228, 237, 332, 335, 337, 340-341
syntactic 48, 288
sub-- 90, 206, 228, 301
change 13, 15-16, 18, 26, 31-32, 51, 63, 67, 72, 74, 109-110, 112-113, 116-117, 128, 132, 149, 151, 266, 279, 354, 368
analagical 67
extra-linguistic 110
lexical 55, 112
regional 110
semantic 13, 15, 51, 110
sociolinguistic 50, 110
sound 67
(morpho-)syntactic 67
see also under meaning
Chinese 61, 294, 299-300, 304-307
Mandarin 63
281, 326, 328, 335, 339-342
because-- 276
causal 274
complement 86, 88-92, 97, 99, 167-168, 171, 177, 180, 182, 185, 188, 190-191, 196, 317, 324, 328-330
dependent 87, 92, 94-95, 97, 104, 274, 318
-ed 191
imperative 274-275, 281
independent 260, 274-276, 279, 281
(to-)infinitive 91, 102, 177, 189, 191
-ing 91
main 90, 101, 260, 263, 267-269, 272-281
matrix 98-99, 259, 322
nou (nominal) 167, 171, 177, 182, 185, 188-191, 196, 316-318, 321, 323, 325, 327, 329
object 96-97, 328
relative 185, 191
subject 85, 97, 100-102, 105, 328
temporal 274, 350-351
that-- 16, 91, 101-102, 173, 177, 189, 328, 333-334, 339-340
wh-- 267, 316, 334
whether-- 167-168, 316-317, 319-330, 334, 338-341
whether/if-- 315-319, 321, 323, 328-334, 336-337, 339-342
see also under adverb(ial), directive, subordination, superordination
coalescence 32
co-event 71-72
cognate 235
cohesion 176, 196
colligation 331, 349, 351, 355
modal 337
semantic 315, 331, 337, 342
textual 165, 347-350, 352, 354-355, 360, 368-369
colocation 25, 43, 331, 340, 349, 351, 355, 368
collostruction 315, 318, 331, 336, 342
command 267, 275-276
declarative 13, 15, 85-86, 88-92, 94-97, 99-100, 102-104
-ed participle 213
factive 100
(non-)finite 86, 88, 91, 212-213
~hood 94
if~ 16, 85-86, 89, 92, 96-105, 329
interrogative 89, 92-95, 99, 168
~izer 13, 15-16, 85-86, 89-92, 94-95, 97-100, 102-104, 276
minor 16, 86, 97, 99, 104
object 185
subject 194
that~ 97, 100-101, 191
though~ 98
topicalized 102-103
[+wh] 85, 95
zero 86, 91, 97, 99
see also under adjective, clause
compositionality 68
condensation 32, 156, 171
conjunction 96, 102, 267, 290
connective 16-18, 34, 85-86, 112-113, 118-119, 121, 124-128, 130, 132
adverbal 13, 16, 109, 112-114, 117-119, 125, 127, 132
[ADJ COMP] 329-334, 337-340
Grammar 15, 68, 318
idiomatic 219-220, 226-227
if~ 16
motion 15, 55, 63, 67, 69-73
pragmatic 279
(non-)prepositional 241-242, 248, 250-251, 253-254
that~ 16
way~ 59, 70, 76
see also get
container 174
negative 334, 336, 340-341
non-assertive 85, 88, 98-100, 315-316, 334, 336, 340-341
see also under meaning
conversion 66, 161-164, 211
data-driven 181, 293-294, 299, 302-304, 309

direction 13, 25, 35, 46-47, 64, 246, 279, 307, 330
divergence 161, 168, 188, 240, 291, 340
domain 62-63, 65, 187, 253, 353

data-driven 181, 293-294, 299, 302-304, 309

post~ 185

domain 62-63, 65, 187, 253, 353

dividend 161, 168, 188, 240, 291, 340
directive 259-262, 264-267, 273, 279-280

directive 259-262, 264-267, 273, 279-280

direction 13, 25, 35, 46-47, 64, 246, 279, 307, 330
divergence 161, 168, 188, 240, 291, 340

domain 62-63, 65, 187, 253, 353

dividend 161, 168, 188, 240, 291, 340
directive 259-262, 264-267, 273, 279-280

direction 13, 25, 35, 46-47, 64, 246, 279, 307, 330
divergence 161, 168, 188, 240, 291, 340

domain 62-63, 65, 187, 253, 353

dividend 161, 168, 188, 240, 291, 340

directive 259-262, 264-267, 273, 279-280

direction 13, 25, 35, 46-47, 64, 246, 279, 307, 330
endophora 176, 185


American 130-131

Asian 206

Australian 261, 264-265, 280

British 96, 130-131, 167, 182, 205-206, 218, 228, 259-260, 266-267, 280-281

contemporary 63-64, 109

Hong Kong 206

Indian 206


Early 118, 141

Late 17, 118-119, 127, 129, 132

Modern 13, 132


Early 111

Late 146

Present-Day 16, 26, 36, 55, 64, 72, 97, 100, 102, 109, 131, 133-134, 141, 143, 145-146, 149-153, 155-156

Scottish 266

Singaporean 206

entropy 148


etymology 28, 30, 235

exophora 176-177, 185, 191

extraposition 100-102, 323-328, 330, 340

field

conceptual 59

lexical 147

semantic 62, 77, 237, 240

fixation 32

Finnish 117, 342

Flemish 65-66

frame 261, 301

P-~ 352, 355, 357, 363-366

satellite-~ 60-63, 66, 247, 254

verb-~ 60-62, 65, 247


Old 65
token 71, 183
type 71, 75
fronting 287, 326, 330
generalization 35, 43, 45, 47, 67, 72, 121, 285, 292, 295-296, 309
German 60-61, 63, 69, 235, 277-278, 285, 287
(Middle) Low 65
gerundive 61
get 167, 205-229
   see also under passive
gradability 210
gradient 85, 93-94, 206-207, 212, 219, 228
   see also under agent
gram 363-364
   bi~ 296-297
   n~ 294-298, 309, 352, 355, 357
   tri~ 296, 298
   uni~ 295
   word-based 295-296, 298
cognitive 67
corpus-based 187
lexico-~ 315, 330-331, 333, 335-336, 340-342, 349, 365
pattern 305, 345, 365
prescriptive 34, 49
systemic-functional 185
traditional 144
see also under construction, interface
de~ 279
granularity 18, 237, 288-289, 292
Hebrew 63, 69
homonymy 145-146
Hungarian 63
hypotheticality 263, 270
Identification
   Native Language 285, 293-294, 298
idiom(atic) 68, 122, 167, 212-213, 224, 239-240, 246
   see also under construction
ilocutionary 259, 265, 270, 273, 279, 281
imperative 262, 267, 276
   see also under clause
implicature 110
informal 85, 102-105, 188
infinitive 276, 361
see also under clause
inflection 144-146, 148, 151, 154, 156, 276, 290
inheritance 68
instrument(ality) 208, 239
intensifier 209
interface
  lexis-grammar 315, 318-319, 342
interjection 76
intonation 101
isolating 156
Japanese 294
jargon 25, 27, 29, 37, 47, 51
keyness 358-359
Latin 60, 65, 113, 117
learn 67, 349, 368
  Computer-Assisted Language
~cographic 77, 162, 164, 166
~colological 162
priming 165-166, 182, 188, 196-197, 347-352, 359, 368
see also under association, category, change, field, grammar, interface
linguistics 13, 67, 178, 287
  applied 161, 164
  cognitive 67, 70, 318
  computational 291, 294
  corpus 55, 57, 318-319, 326
  descriptive 161, 164, 169
  historical 13, 26, 112
  socio~  18, 44, 49, 51, 110-111, 132, 292, 298
loan 65, 117, 119, 132
  see also under alternation
manner 14-15, 55-56, 60-67, 69, 71, 76, 244-248, 252, 269
  see also under motion, verb
matrix 87, 95, 97-100
  affirmative 98
  (non-)assertive 98-100, 105
  interrogative 98
  negated 98
  superordinate 259
  see also under clause
abstract 27, 30-32, 35, 37, 48, 50
change 25-27, 35, 50-51
component 60, 234
collection-dependent 175, 189
dictionary 173, 175
dynamic 207
extension 27, 35, 50
figurative 27, 35, 49
literal 25, 29, 35, 41, 49, 168
metaphorical 27, 29-30, 35, 37, 40, 44, 46-47, 49
stative 207, 210
medium 98, 239-240, 266
mental, 62, 186, 189, 194-195, 205, 223, 228, 349
metaphor 14, 25, 27, 29-31, 33, 35-37, 39, 44-49, 74, 234, 240, 251-254
see also under meaning
metonymy 35
see also under construction, verb
modification 209, 262, 342
post- 177, 185, 190-191, 193, 196, 245
pre- 180, 211, 361
mood 342
morpheme 32
morphology 18, 32, 35, 112, 141, 143, 145, 163, 211, 269, 279, 288, 290-291, 299, 334-335, 337, 341
mark 141, 144, 148, 290
manner 13, 15, 55, 62-67, 76, 244-246
see also under construction, verb
negation 89, 97-98, 168, 189, 262, 277-278, 315-316, 319, 333-337, 340-342
see also under context, matrix, phrase
Norse
Old 30, 65, 113
abstract 171-172, 195, 213
nominalization 174
shell 166-167, 171-184, 186-197
see also under clause, phrase
object 96-97, 122-123, 194, 213, 287, 299, 302, 304, 316, 323, 327, 340
direct 185, 194, 197, 307
prepositional 299, 301
see also under clause, complement
obligatory 32-33, 90-92, 94
onomasiology 235, 237
optional 61, 117, 263, 287, 298-299, 301-305
paradigm(atic/atization) 18, 32, 144-146, 149
parse(r/ing) 116, 218, 267, 300-301
participant 185, 194, 196, 218, 262-263, 266-268, 272
participle 276
-ed-(past) 191, 206-207, 209-211, 213, 219, 223, 228

see also under adjective, complement
passive 167, 206-208, 210, 213-215, 217, 219, 221-222, 328
adjectival 167, 211, 219-220, 226-228
agentless 208-209, 214, 221
be~~ 167, 205, 214-217, 223, 229
get~~ 167, 205-229
pseudo~~ 167, 210, 213, 219-220, 224, 226-227
reflexive (reciprocal) 167, 210, 212-213, 219, 221, 224-228
semi~~ 167, 209, 213, 219-220, 222, 224, 226-229
path 60-61, 63-64, 71-73, 233, 238-239, 244-248, 252-254

see also under verb
perception 169, 237-238, 240, 247-249, 253
phonaesteme 75
phonetic 67
motivation 73
symbolism 74
adjective 97, 105, 323, 328, 330, 340
adverb 61, 287
agent 207-211, 213-215, 221-222, 227-228
interrogative 89
negative 89
noun 34, 71-72, 97, 105, 113, 172, 175-177, 181, 184-186, 188-189, 193, 212-213, 296, 300-302, 304-307
of~~ 190, 193
verb 98, 100, 244-248, 254, 269, 363
with~~ 208
polarity 87-88, 315, 331, 336-339, 342
polysemy 28-29, 44
pragmatics 32, 167, 171-172, 205, 209, 228, 235, 262, 269-270, 278-279, 293, 342, 349

see also under annotation, association, construction
predicate 86, 90, 92-99, 168, 211, 238-239, 248, 252-253, 276, 324-328, 330, 340-341
commentative 85, 99, 105
spatial 238
temporal 238
prefix 13, 27, 29-31, 34-36, 47, 49-50, 335

see also under verb
preposition 16-17, 27, 31-34, 36, 112-114, 117, 122-123, 125-126, 128-129, 132, 168, 190, 194, 207-208, 210, 222, 234-236, 241-250, 253, 290, 300
~/particle 168, 237
(in-)transitive 34

see also under alternation, construction, object, phrase
productivity 31, 68, 70, 163
pronoun 91-92, 94, 100-102, 105, 117, 175-176, 336-338

proposition 90, 97-98, 117, 124, 174, 244, 268, 270, 348, 351, 360-363
pseudo-clefting 91-92, 94
punctuation 101, 275, 295, 357

question 87-88, 93, 104, 316, 333, 337-342

reduction 31-32, 272

see also anaphora, cataphora, endophora, exophora
register 16-18, 77, 104, 130, 167, 191
speech-based 16, 103-105
written 103-104
request 259-260, 262-266, 268, 270-271, 273, 276-277, 279-280
responsibility 167, 205, 208, 212, 214, 216-217, 225-229
rheme 185, 194-195, 197

Romanian 342
Russian 61, 63, 294, 342

satellite 60-61

see also under frame
scale 93-95, 174, 205, 208-211, 216

Scandinavian 66


extension 31, 66, 166
preference 331
specificity 62-63

see also under annotation, argument, association, category, change, colligation, domain, field

semasiology 235

Semitic 61

sense 27, 29, 35, 37, 45-46, 69, 72, 74, 89, 119, 121-122, 162, 169, 233, 238-240, 243-244, 248, 252-254, 288, 326, 333


~boundary 176, 185, 356

paragraph-final 165, 352, 355, 357-359, 362, 364, 366, 368-369
paragraph-initial 165, 352, 355, 357-362, 364-368
paragraph-medial 352, 357-359, 362, 364
space 27, 35, 45, 233-234, 237-240, 245

see also under predicate
Spanish 61, 69, 117
speech 16, 34, 103-105, 189, 276
act 273
spelling 38, 40, 115, 147
standard 40
variation 115-116, 145-147
subject 63, 85, 96-97, 100-102, 104, 119-120, 122-123, 167, 185, 194, 197, 205-206, 208, 211, 213-218, 222, 224-229, 248, 250, 252, 265, 287, 300, 323-326, 328, 330, 339-341, 360

see also under clause, complement
(in)subordination 102, 259-260, 276-279, 281, 317
adverbial 13, 85, 113, 117
clause 86, 90, 268, 272, 274, 276-279, 281, 315-317, 323, 333-334, 337, 341
conditional 87
subordinator 85, 87-89, 104, 112-114, 117, 124, 318-319, 338, 342
superordination
clause 328

see also under matrix
stem 145, 148, 175, 178, 290-291
suffix 73, 290
Swedish 61, 69, 117
synonymy 59, 72
syntheticity 13, 18-19, 141-144, 148-156
grammatical 141-142

index 18, 142, 144, 148-149, 151-153
tag(ger/ging) 18, 116, 144, 151, 218, 289-291, 295, 301, 318
teaching 286-287, 289
tertium comparationis 233-237

see also co-text

see also under colligation, coordination, domain

theme 185, 194-195, 197, 350, 359
thesaurus 15, 59, 64, 166
'throughness' 168-169, 233-235, 237-240, 242-244, 246-249, 251-254
time 29, 33, 35, 37, 39-40, 44-45, 47-48, 112-113, 239-240, 249, 253, 265, 269, 281, 351

see also under clause, predicate
token(ization) 17, 142, 144-146, 148-149, 219, 221-222, 233, 236-251, 253-254, 267, 295, 332
see also under frequency
trajector 233, 239
translation 62-63, 65, 73, 117, 233-245, 250, 252-253
(in)transitivity 15, 34, 63, 65, 70-73, 185, 195, 302
catenative 212
transformative 213
see also under preposition
Turkish 61, 63
typology 13, 18, 61, 141-142, 153, 156, 169, 247, 254
quantitative 156

undergraduate 178, 180-181, 300, 347-349, 352, 368
usage 15, 18, 25, 27, 34-35, 43-46, 49-51, 57-58, 67-68, 131, 163, 239, 264

see also under spelling
variety 18, 39, 77, 111, 117, 129, 143, 148, 167, 206, 291
attributive 328

auxiliary 210-211, 213, 270, 287, 358, 363
catenative 212
causative 213, 223
copula(r) 97, 105, 209-211
dynamic 214-215, 223
frequentative 73
~hood 76
~less 267
manner 13, 55, 61-66, 71, 76, 244-246
modal 98, 165, 261, 265, 269-271, 359, 367-368
motion 13-15, 25, 63-65, 71, 244-246, 306
multi-word 239
non-agentive 63
particle 60-61
path 61-62, 64, 245-246
prefix 60-61
reational 324
sound emission 15, 69-70, 72, 75
stative 215
see also (in)transitivity
see also under frame, phrase
vocabulary 26, 55, 59, 112-113, 124, 174
volition 208, 216, 270

want to/wanna 271
will 270-271, 358, 368
word(-)class 18, 33, 114, 144-146, 148-149, 151-154, 218, 291
word formation 31-32, 66
would 98, 100, 270-271, 280, 358-359, 368
Linguistic Insights
Studies in Language and Communication

This series aims to promote specialist language studies in the fields of linguistic theory and applied linguistics, by publishing volumes that focus on specific aspects of language use in one or several languages and provide valuable insights into language and communication research. A cross-disciplinary approach is favoured and most European languages are accepted.

The series includes two types of books:

- **Monographs** – featuring in-depth studies on special aspects of language theory, language analysis or language teaching.
- **Collected papers** – assembling papers from workshops, conferences or symposia.

Each volume of the series is subjected to a double peer-reviewing process.

Vol. 1  Maurizio Gotti & Marina Dossena (eds)
Modality in Specialized Texts. Selected Papers of the 1st CERLIS Conference.

Vol. 2  Giuseppina Cortese & Philip Riley (eds)
Domain-specific English. Textual Practices across Communities and Classrooms.

Vol. 3  Maurizio Gotti, Dorothee Heller & Marina Dossena (eds)
Conflict and Negotiation in Specialized Texts. Selected Papers of the 2nd CERLIS Conference.

Vol. 4  Maurizio Gotti, Marina Dossena, Richard Dury, Roberta Facchinetti & Maria Lima
Variation in Central Modals. A Repertoire of Forms and Types of Usage in Middle English and Early Modern English.

**Editorial address:**

Prof. Maurizio Gotti  Università di BergamoDipartimento di Lingue, Letterature Straniere e Comunicazione, Piazza Rosate 2, 24129 Bergamo, Italy
Fax: +39 035 2052789, E-Mail: m.gotti@unibg.it
Vol. 5  Stefania Nuccorini (ed.)
Phrases and Phraseology. Data and Descriptions.

Vol. 6  Vijay Bhatia, Christopher N. Candlin & Maurizio Gotti (eds)
Legal Discourse in Multilingual and Multicultural Contexts.
Arbitration Texts in Europe.

Vol. 7  Marina Dossena & Charles Jones (eds)
Insights into Late Modern English. 2nd edition.

Vol. 8  Maurizio Gotti
Specialized Discourse. Linguistic Features and Changing Conventions.

Vol. 9  Alan Partington, John Morley & Louann Haarman (eds)
Corpora and Discourse.

Vol. 10  Martina Möllering
The Acquisition of German Modal Particles. A Corpus-Based Approach.

Vol. 11  David Hart (ed.)
English Modality in Context. Diachronic Perspectives.

Vol. 12  Wendy Swanson
Modes of Co-reference as an Indicator of Genre.

Vol. 13  Gina Poncini
Discursive Strategies in Multicultural Business Meetings.

Vol. 14  Christopher N. Candlin & Maurizio Gotti (eds)
Intercultural Aspects of Specialized Communication.

Vol. 15  Gabriella Del Lungo Camiciotti & Elena Tognini Bonelli (eds)

Vol. 16  Marina Dossena & Roger Lass (eds)
Methods and Data in English Historical Dialectology.

Vol. 17  Judy Noguchi
The Science Review Article. An Opportune Genre in the Construction of Science.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>32</td>
<td>Business and Official Correspondence. Historical Investigations.</td>
<td>Marina Dossena &amp; Susan M. Fitzmaurice</td>
<td>209</td>
</tr>
<tr>
<td>33</td>
<td>Discourse, Ideology and Specialized Communication.</td>
<td>Giuliana Garzone &amp; Srikant Sarangi</td>
<td>494</td>
</tr>
<tr>
<td>34</td>
<td>The Use of English in Institutional and Business Settings. An Intercultural Perspective.</td>
<td>Giuliana Garzone &amp; Cornelia Ilie</td>
<td>372</td>
</tr>
<tr>
<td>36</td>
<td>Modern Approaches to Terminological Theories and Applications.</td>
<td>Heribert Picht (ed.)</td>
<td>432</td>
</tr>
<tr>
<td>37</td>
<td>Legal Language and the Search for Clarity / Le langage juridique et la quête de clarté. Practice and Tools / Pratiques et instruments.</td>
<td>Anne Wagner &amp; Sophie Cacciaguidi-Fahy (eds)</td>
<td>487</td>
</tr>
<tr>
<td>40</td>
<td>Diachronic Perspectives on Domain-Specific English.</td>
<td>Marina Dossena &amp; Irma Taavitsainen (eds)</td>
<td>280</td>
</tr>
<tr>
<td>41</td>
<td>Studies in Specialized Discourse.</td>
<td>John Flowerdew &amp; Maurizio Gotti (eds)</td>
<td>293</td>
</tr>
<tr>
<td>42</td>
<td>Academic Discourse Across Disciplines.</td>
<td>Ken Hyland &amp; Marina Bondi (eds)</td>
<td>320</td>
</tr>
<tr>
<td>43</td>
<td>The Map and the Landscape. Norms and Practises in Genre.</td>
<td>Paul Gillaerts &amp; Philip Shaw (eds)</td>
<td>256</td>
</tr>
<tr>
<td>45</td>
<td>Advances in Medical Discourse Analysis. Oral and Written Contexts.</td>
<td>Maurizio Gotti &amp; Françoise Salager-Meyer (eds)</td>
<td>492</td>
</tr>
</tbody>
</table>
Vol. 61  Sally Burgess & Pedro Martín-Martín (eds)
English as an Additional Language in Research Publication and Communication.

Vol. 62  Sandrine Onillon
Pratiques et représentations de l’écrit.

Vol. 63  Hugo Bowles & Paul Seedhouse (eds)
Conversation Analysis and Language for Specific Purposes.

Vol. 64  Vijay K. Bhatia, Christopher N. Candlin & Paola Evangelisti Allori (eds)
Language, Culture and the Law.
The Formulation of Legal Concepts across Systems and Cultures.

Vol. 65  Jonathan Culpeper & Dániel Z. Kádár (eds)
Historical (Im)politeness.

Vol. 66  Linda Lombardo (ed.)
Using Corpora to Learn about Language and Discourse.

Vol. 67  Natsumi Wakamot
Extraversion/Introversion in Foreign Language Learning.
Interactions with Learner Strategy Use.

Vol. 68  Eva Alcón-Soler (ed.)
Learning How to Request in an Instructed Language Learning Context.

Vol. 69  Domenico Pezzini
The Translation of Religious Texts in the Middle Ages.

Vol. 70  Tomoko Tode
Effects of Frequency in Classroom Second Language Learning.
Quasi-experiment and stimulated-recall analysis.

Vol. 71  Egor Tsedryk
Fusion symétrique et alternances ditransitives.

Vol. 72  Cynthia J. Kellett Bidoli & Elana Ochse (eds)
English in International Deaf Communication.

Vol. 73  Joan C. Beal, Carmela Nacera & Massimo Sturiale (eds)
Perspectives on Prescriptivism.

Vol. 74  Carol Taylor Torsello, Katherine Ackerley & Erik Castello (eds)
Corpora for University Language Teachers.
Vol. 75 María Luisa Pérez Cañado (ed.)
English Language Teaching in the European Credit Transfer System. Facing the Challenge.

Vol. 76 Marina Dossena & Ingrid Tieken-Boon van Ostade (eds)
Studies in Late Modern English Correspondence. Methodology and Data.

Vol. 77 Ingrid Tieken-Boon van Ostade & Wim van der Wurff (eds)
Current Issues in Late Modern English.

Vol. 78 Marta Navarro Coy (ed.)
Practical Approaches to Foreign Language Teaching and Learning.
297 pages. 2009. ISBN 978-3-03911-661-4

Vol. 79 Qing Ma
Second Language Vocabulary Acquisition.

Vol. 80 Martin Solly, Michelangelo Conoscenti & Sandra Campagna (eds)
Verbal/Visual Narrative Texts in Higher Education.

Vol. 81 Meiko Matsumoto
From Simple Verbs to Periphrastic Expressions: The Historical Development of Composite Predicates, Phrasal Verbs, and Related Constructions in English.

Vol. 82 Melinda Dooly
Doing Diversity. Teachers’ Construction of Their Classroom Reality.

Vol. 83 Victoria Guillén-Nieto, Carmen Marimón-Llorca & Chelo Vargas-Sierra (eds)
Intercultural Business Communication and Simulation and Gaming Methodology.

Vol. 84 Maria Grazia Guido
English as a Lingua Franca in Cross-cultural Immigration Domains.

Vol. 85 Erik Castello
Text Complexity and Reading Comprehension Tests.

Vol. 86 Maria-Lluisa Gea-Valor, Isabel García-Izquierdo & Maria-José Esteve (eds)

Vol. 87 Carmen Navarro, Rosa Mª Rodríguez Abella, Francesca Dalle Pezze & Renzo Miotti (eds)
La comunicación especializada.
<table>
<thead>
<tr>
<th>Vol.</th>
<th>Title</th>
<th>Authors/Editors</th>
<th>Pages</th>
<th>ISBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>The Development from Case-Forms to Prepositional Constructions in Old English Prose.</td>
<td>Kiriko Sato</td>
<td>231</td>
<td>978-3-03911-763-5</td>
</tr>
<tr>
<td>89</td>
<td>Formulierungsmuster in deutscher und italienischer Fachkommunikation. Intra- und interlinguale Perspektiven.</td>
<td>Dorothee Heller (Hrsg.)</td>
<td>315</td>
<td>978-3-03911-778-9</td>
</tr>
<tr>
<td>91</td>
<td>The Development of Controversies. From the Early Modern Period to Online Discussion Forums.</td>
<td>Manouchehr Moshtagh Khorasani</td>
<td>317</td>
<td>978-3-3911-711-6</td>
</tr>
<tr>
<td>93</td>
<td>Terminology in English Language Teaching. Nature and Use.</td>
<td>Roger Berry</td>
<td>262</td>
<td>978-3-0343-0013-1</td>
</tr>
<tr>
<td>94</td>
<td>Discourses, Communities, and Global Englishes</td>
<td>Roberto Caglieri &amp; Jennifer Jenkins (eds)</td>
<td>240</td>
<td>978-3-0343-0012-4</td>
</tr>
<tr>
<td>95</td>
<td>From International to Local English – And Back Again.</td>
<td>Facchinetti Roberta, Crystal David, Seidlhofer Barbara (eds)</td>
<td>268</td>
<td>978-3-0343-0011-7</td>
</tr>
<tr>
<td>96</td>
<td>EIL, ELF, Global English. Teaching and Learning Issues</td>
<td>Cesare Gagliardi &amp; Alan Maley (eds)</td>
<td>376</td>
<td>978-3-0343-0010-0</td>
</tr>
<tr>
<td>97</td>
<td>The Role of Prosody in Affective Speech.</td>
<td>Sylvie Hancil (ed.)</td>
<td>403</td>
<td>978-3-03911-696-6</td>
</tr>
<tr>
<td>98</td>
<td>Studies in English and European Historical Dialectology.</td>
<td>Marina Dossena &amp; Roger Lass (eds)</td>
<td>257</td>
<td>978-3-0343-0024-7</td>
</tr>
<tr>
<td>99</td>
<td>Les interactions quotidiennes en français et en anglais. De l’approche comparative à l’analyse des situations interculturelles.</td>
<td>Christine Béal</td>
<td>424</td>
<td>978-3-0343-0027-8</td>
</tr>
<tr>
<td>100</td>
<td>Commonality and Individuality in Academic Discourse.</td>
<td>Maurizio Gotti (ed.)</td>
<td>398</td>
<td>978-3-0343-0023-0</td>
</tr>
<tr>
<td>Vol.</td>
<td>Authors</td>
<td>Title</td>
<td>Pages</td>
<td>ISBN</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------</td>
<td>--------------------</td>
</tr>
<tr>
<td>102</td>
<td>Nuria Edo Marzá</td>
<td>The Specialised Lexicographical Approach. A Step further in Dictionary-making.</td>
<td>316</td>
<td>978-3-0343-0043-8</td>
</tr>
<tr>
<td>105</td>
<td>Javier Ruano-García</td>
<td>Early Modern Northern English Lexis. A Literary Corpus-Based Study.</td>
<td>611</td>
<td>978-3-0343-0058-2</td>
</tr>
<tr>
<td>106</td>
<td>Rafael Monroy-Casas</td>
<td>Systems for the Phonetic Transcription of English. Theory and Texts.</td>
<td>280</td>
<td>978-3-0343-0059-9</td>
</tr>
<tr>
<td>107</td>
<td>Nicola T. Owtram</td>
<td>The Pragmatics of Academic Writing. A Relevance Approach to the Analysis of Research Article Introductions.</td>
<td>311</td>
<td>978-3-0343-0060-5</td>
</tr>
<tr>
<td>110</td>
<td>Rosalía Rodríguez-Vázquez</td>
<td>The Rhythm of Speech, Verse and Vocal Music. A New Theory.</td>
<td>394</td>
<td>978-3-0343-0309-5</td>
</tr>
<tr>
<td>111</td>
<td>Anastasios Tsangalidis &amp; Roberta Focchini (eds)</td>
<td>Studies on English Modality. In Honour of Frank Palmer.</td>
<td>392</td>
<td>978-3-0343-0310-1</td>
</tr>
<tr>
<td>112</td>
<td>Jing Huang</td>
<td>Autonomy, Agency and Identity in Foreign Language Learning and Teaching.</td>
<td>400</td>
<td>978-3-0343-0370-5</td>
</tr>
<tr>
<td>113</td>
<td>Mihhail Lotman &amp; Maria-Kristiina Lotman (eds)</td>
<td>Frontiers in Comparative Prosody. In memoriam: Mikhail Gasparov.</td>
<td>426</td>
<td>978-3-0343-0373-6</td>
</tr>
<tr>
<td>114</td>
<td>Merja Kytö, John Scahill &amp; Harumi Tanabe (eds)</td>
<td>Language Change and Variation from Old English to Late Modern English. A Festschrift for Minoji Akimoto</td>
<td>422</td>
<td>978-3-0343-0372-9</td>
</tr>
<tr>
<td>115</td>
<td>Giuliana Garzone &amp; Paola Catenaccio (eds)</td>
<td>Identities across Media and Modes. Discursive Perspectives.</td>
<td>379</td>
<td>978-3-0343-0386-6</td>
</tr>
<tr>
<td>Vol.</td>
<td>Title</td>
<td>Author(s)</td>
<td>ISBN</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>Los marcadores del discurso y cortesía verbal en español.</td>
<td>Elena Landone</td>
<td>978-3-0343-0413-9</td>
<td></td>
</tr>
<tr>
<td>117</td>
<td>Legal Discourse across Languages and Cultures.</td>
<td>Maurizio Gotti &amp; Christopher Williams (eds)</td>
<td>978-3-0343-0425-2</td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>Academic Vocabulary in Context.</td>
<td>David Hirsh</td>
<td>978-3-0343-0426-9</td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>Lingua Franca English. The Role of Simplification and Transfer.</td>
<td>Yvonne Dröschel</td>
<td>978-3-0343-0432-0</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>Corpora in Translation. A Practical Guide.</td>
<td>Tengku Sepora Tengku Mahadi, Helia Vaezian &amp; Mahmoud Akbari</td>
<td>978-3-0343-0434-4</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>Researching Language and the Law. Textual Features and Translation Issues.</td>
<td>Davide Simone Giannoni &amp; Celina Frade (eds)</td>
<td>978-3-0343-0443-6</td>
<td></td>
</tr>
<tr>
<td>122</td>
<td>Studies in Bilingual Education.</td>
<td>Daniel Madrid &amp; Stephen Hughes (eds)</td>
<td>978-3-0343-0474-0</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>The Discourses of Dispute Resolution.</td>
<td>Vijay K. Bhatia, Christopher N. Candlin &amp; Maurizio Gotti (eds)</td>
<td>978-3-0343-0476-4</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>Mapping Academic Values in the Disciplines. A Corpus-Based Approach.</td>
<td>Davide Simone Giannoni</td>
<td>978-3-0343-0488-7</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>Discourse, Identities and Roles in Specialized Communication.</td>
<td>Giuliana Garzone &amp; James Archibald (eds)</td>
<td>978-3-0343-0494-8</td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>The Status and Development of N+N Sequences in Contemporary English Noun Phrases.</td>
<td>Iria Pastor-Gómez</td>
<td>978-3-0343-0534-1</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>Full-verb Inversion in Written and Spoken English.</td>
<td>Carlos Prado-Alonso</td>
<td>978-3-0343-0535-8</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>Corpus Linguistics in Language Teaching.</td>
<td>Tony Harris &amp; María Moreno Jaén (eds)</td>
<td>978-3-0343-0524-2</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>Multiple Perspectives on English Philology and History of Linguistics. A Festschrift for Shoichi Watanabe on his 80th Birthday.</td>
<td>Tetsuji Oda &amp; Hiroyuki Eto (eds)</td>
<td>978-3-0343-0480-1</td>
<td></td>
</tr>
</tbody>
</table>
Vol. 130  Luisa Chierichetti & Giovanni Garofalo (eds)  
Lengua y Derecho. líneas de investigación interdisciplinaria.  
283 pages. 2010. 978-3-0343-0463-4

Vol. 131  Paola Evangelisti Allori & Giuliana Garzone (eds)  
Discourse, Identities and Genres in Corporate Communication.  
Sponsorship, Advertising and Organizational Communication.  
324 pages. 2011. 978-3-0343-0591-4

Vol. 132  Leyre Ruiz de Zarobe & Yolanda Ruiz de Zarobe (eds)  
Speech Acts and Politeness across Languages and Cultures.  
402 pages. 2012. 978-3-0343-0611-9

Vol. 133  Thomas Christiansen  
Cohesion. A Discourse Perspective.  
387 pages. 2011. 978-3-0343-0619-5

Vol. 134  Giuliana Garzone & Maurizio Gotti  

Vol. 135  Zsuzsa Hoffmann  
Ways of the World’s Words. Language Contact in the Age of Globalization.  

Vol. 136  Cecilia Varcasia (ed.)  
Becoming Multilingual. Language Learning and Language Policy between Attitudes and Identities.  

Vol. 137  Susy Macqueen  
The Emergence of Patterns in Second Language Writing. A Sociocognitive Exploration of Lexical Trails.  

Vol. 138  Maria Vittoria Calvi & Giovanna Mapelli (eds)  
La lengua del turismo. Géneros discursivos y terminología.  

Vol. 139  Ken Lau  

Vol. 140  Sandra Campagna, Giuliana Garzone, Cornelia Ilie & Elizabeth Rowley-Jolivet (eds)  
Evolving Genres in Web-mediated Communication.  

Vol. 141  Edith Esch & Martin Solly (eds)  
The Sociolinguistics of Language Education in International Contexts.  

Vol. 142  Forthcoming.

Vol. 143  David Tizón-Couto  
Left Dislocation in English. A Functional-Discoursal Approach.  
Vol. 144 Margrethe Petersen & Jan Engberg (eds)  

Vol. 145 David Tizón-Couto, Beatriz Tizón-Couto, Iria Pastor-Gómez & Paula Rodríguez-Puente (eds)  
New Trends and Methodologies in Applied English Language Research II.  
Studies in Language Variation, Meaning and Learning.  
283 pages. 2012. ISBN 978-3-0343-1061-1

Vol. 146 Rita Salvi & Hiromasa Tanaka (eds)  
Intercultural Interactions in Business and Management.  

Vol. 147 Francesco Straniero Sergio & Caterina Falbo (eds)  
Breaking Ground in Corpus-based Interpreting Studies.  

Vol. 148 Forthcoming.

Vol. 149 Vijay K. Bhatia & Paola Evangelisti Allori (eds)  
Discourse and Identity in the Professions. Legal, Corporate and Institutional Citizenship.  

Vol. 150 Maurizio Gotti (ed.)  
Academic Identity Traits. A Corpus-Based Investigation.  

Vol. 151 Priscilla Heynderickx, Sylvain Dieltjens, Geert Jacobs, Paul Gillaerts & Elizabeth de Groot (eds)  
The Language Factor in International Business.  
New Perspectives on Research, Teaching and Practice.  

Vol. 152 Paul Gillaerts, Elizabeth de Groot, Sylvain Dieltjens, Priscilla Heynderickx & Geert Jacobs (eds)  

Vol. 153 Yongyan Zheng  
Dynamic Vocabulary Development in a Foreign Language.  

Vol. 154 Carmen Argondizzo (ed.)  
Creativity and Innovation in Language Education.  

Vol. 155 David Hirsh (ed.)  
Current Perspectives in Second Language Vocabulary Research.  

Vol. 156 Seiji Shinkawa  
Unhistorical Gender Assignment in Laṣamōn’s Brut. A Case Study of a Late Stage in the Development of Grammatical Gender toward its Ultimate Loss.  

Vol. 157 Yeonkwon Jung  
<table>
<thead>
<tr>
<th>Vol.</th>
<th>Title</th>
<th>Editors/Authors</th>
<th>Pages</th>
<th>ISBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>159</td>
<td>Genre Change in the Contemporary World. Short-term Diachronic Perspectives.</td>
<td>Giuliana Garzone, Paola Catenaccio, Chiara Degano (eds)</td>
<td>329</td>
<td>978-3-0343-1214-1</td>
</tr>
<tr>
<td>160</td>
<td>Insights into Academic Genres.</td>
<td>Carol Berkenkotter, Vijay K. Bhatia &amp; Maurizio Gotti (eds)</td>
<td>468</td>
<td>978-3-0343-1211-0</td>
</tr>
<tr>
<td>161</td>
<td>Clausal Complements in Native and Learner Spoken English. A corpus-based study with Lindsei and Vicolse.</td>
<td>Beatriz Tizón-Cauto</td>
<td>357</td>
<td>978-3-0343-1184-7</td>
</tr>
<tr>
<td>162</td>
<td>Jury Trials and the Popularization of Legal Language. A Discourse Analytical Approach.</td>
<td>Patrizia Anesa</td>
<td>247</td>
<td>978-3-0343-1231-8</td>
</tr>
<tr>
<td>163</td>
<td>Endangered Languages, Knowledge Systems and Belief Systems.</td>
<td>David Hirsh</td>
<td>153</td>
<td>978-3-0343-1232-5</td>
</tr>
<tr>
<td>164</td>
<td>De la estructura de la frase al tejido del discurso. Estudios contrastivos español/italiano.</td>
<td>Eugenia Sainz (ed.)</td>
<td>305</td>
<td>978-3-0343-1253-0</td>
</tr>
<tr>
<td>165</td>
<td>Space, Place and the Discursive Construction of Identity.</td>
<td>Julia Bamford, Franca Poppi &amp; Davide Mazzi (eds)</td>
<td>367</td>
<td>978-3-0343-1249-3</td>
</tr>
<tr>
<td>166</td>
<td>Discourse Indexicality in Cultural, Institutional and Professional Fields.</td>
<td>Rita Salvi &amp; Janet Bowker (eds)</td>
<td>324</td>
<td>978-3-0343-1254-7</td>
</tr>
<tr>
<td>167</td>
<td>Approaching Language Variation through Corpora. A Festschrift in Honour of Toshio Saito.</td>
<td>Shunji Yamazaki &amp; Robert Sigley (eds)</td>
<td>421</td>
<td>978-3-0343-1264-6</td>
</tr>
<tr>
<td>168</td>
<td>Global Interactions in English as a Lingua Franca. How written communication is changing</td>
<td>Franca Poppi</td>
<td>249</td>
<td>978-3-0343-1276-9</td>
</tr>
<tr>
<td>169</td>
<td>Style in syntax. Investigating variation in Spanish pronoun subjects.</td>
<td>Miguel A. Aijón Oliva &amp; María José Serrano</td>
<td>239</td>
<td>978-3-0343-1244-8</td>
</tr>
<tr>
<td>170</td>
<td>Language Use in the Public Sphere. Methodological Perspectives and Empirical Applications</td>
<td>Manuel Casado, Óscar Loureda &amp; Inés Olza (eds)</td>
<td>564</td>
<td>978-3-0343-1286-8</td>
</tr>
<tr>
<td>172</td>
<td>Narratives in Academic and Professional Genres.</td>
<td>Maurizio Gotti &amp; Carmen Sancho Guinda (eds)</td>
<td>513</td>
<td>978-3-0343-1371-1</td>
</tr>
<tr>
<td>Vol.</td>
<td>Title</td>
<td>Author(s)</td>
<td>ISBN</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td>173</td>
<td>Forthcoming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>Task Equivalence in Speaking Tests.</td>
<td>Chihiro Inoue</td>
<td>978-3-0343-1417-6</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>LSP in Colombia: advances and challenges.</td>
<td>Gabriel Quiroz &amp; Pedro Patiño (eds.)</td>
<td>978-3-0343-1434-3</td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>Forthcoming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>178</td>
<td>Le discours rapporté: approches linguistiques et perspectives didactiques.</td>
<td>Cécile Desoutter &amp; Caroline Mellet (dir.)</td>
<td>978-3-0343-1292-9</td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>A Multi-dimensional Approach to Discourse Coherence. From Standardness to Creativity.</td>
<td>Pilar Alonso</td>
<td>978-3-0343-1325-4</td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>Diachrony and Synchrony in English Corpus Linguistics.</td>
<td>Alejandro Alcaraz-Sintes &amp; Salvador Valera-Hernández (eds)</td>
<td>978-3-0343-1326-1</td>
<td></td>
</tr>
<tr>
<td>182</td>
<td>Forthcoming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>183</td>
<td>English in Malaysia. Postcolonial and Beyond.</td>
<td>Hajar Abdul Rahim &amp; Shakila Abdul Manan (eds)</td>
<td>978-3-0343-1341-4</td>
<td></td>
</tr>
<tr>
<td>184</td>
<td>Comprendre et apprendre dans l’interaction. Les séquences d’explication en classe de français langue seconde.</td>
<td>Virginie Fasel Lauzon</td>
<td>978-3-0343-1451-0</td>
<td></td>
</tr>
<tr>
<td>185-186</td>
<td>Forthcoming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>187</td>
<td>Abstracts in Academic Discourse. Variation and Change.</td>
<td>Marina Bondi &amp; Rosa Lorés Sanz (eds)</td>
<td>978-3-0343-1483-1</td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>Forthcoming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>189</td>
<td>Identities in and across Cultures.</td>
<td>Paola Evangelisti Allori (ed.)</td>
<td>978-3-0343-1458-9</td>
<td></td>
</tr>
<tr>
<td>190</td>
<td>Forthcoming</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>191</td>
<td>Interpersonality in Legal Genres.</td>
<td>Ruth Breeze, Maurizio Gotti &amp; Carmen Sancho Guinda (eds)</td>
<td>978-3-0343-1524-1</td>
<td></td>
</tr>
</tbody>
</table>
Vol. 192-199 Forthcoming.

Vol. 200  Maurizio Gotti & Davide S. Giannoni (eds)
Corpus Analysis for Descriptive and Pedagogical Purposes. ESP Perspectives.
432 pages. 2014. ISBN 978-3-0343-1516-6