New Approaches to English Linguistics

Building bridges

Edited by
Olga Timofeeva
Anne-Christine Gardner
Alpo Honkapohja
Sarah Chevalier
New Approaches to English Linguistics
Studies in Language Companion Series (SLCS)

ISSN 0165-7763

This series has been established as a companion series to the periodical Studies in Language.

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

Founding Editor

Werner Abraham
University of Vienna / University of Munich

Editors

Werner Abraham
University of Vienna / University of Munich

Elly van Gelderen
Arizona State University

Editorial Board

Bernard Comrie
University of California, Santa Barbara

Christian Lehmann
University of Erfurt

William Croft
University of New Mexico

Marianne Mithun
University of California, Santa Barbara

Östen Dahl
University of Stockholm

Heiko Narrog
Tohoku University

Gerrit J. Dimmendaal
University of Cologne

Johanna L. Wood
University of Aarhus

Ekkehard König
Free University of Berlin

Debra Ziegeler
University of Paris III

Volume 177

New Approaches to English Linguistics. Building bridges
Edited by Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja and Sarah Chevalier
New Approaches to English Linguistics
Building bridges

Edited by
Olga Timofeeva
Anne-Christine Gardner
Alpo Honkapohja
Sarah Chevalier
University of Zurich

John Benjamins Publishing Company
Amsterdam/Philadelphia
# Table of contents

Introduction  
*Sarah Chevalier, Anne-Christine Gardner, Alpo Honkapohja, Marianne Hundt, Gerold Schneider and Olga Timofeeva*

Chapter 1  
Accommodation, dialect contact and grammatical variation: Verbs of obligation in the Anglophone community in Japan  
*Keiko Hirano & David Britain*

Chapter 2  
Patterns of linguistic globalization: Integrating typological profiles and questionnaire data  
*Manfred Krug, Ole Schützler & Valentin Werner*

Chapter 3  
The substitutability and diffusion of *want to* and *wanna* in world Englishes  
*Eugene Green*

Chapter 4  
Dialect contact influences on the use of *get* and the *get-passive*  
*Elisabeth Bruckmaier*

Chapter 5  
Future time marking in spoken Ghanaian English: The variation of *will* vs. *be going to*  
*Agnes Schneider*

Chapter 6  
Ongoing changes in English modals: On the developments in ELF  
*Mikko Laitinen*
Chapter 7
Building interdisciplinary bridges: MUCH: The Malmö University-Chalmers Corpus of Academic Writing as a Process
Anna Wärnsby, Asko Kauppinen, Andreas Eriksson, Maria Wiktorsson, Eckhard Bick & Leif-Jöran Olsson

Chapter 8
Discourse markers in L2 English: From classroom to naturalistic input
Gaëtanelle Gilquin

Chapter 9
Processing of aspectual meanings by non-native and native English speakers during narrative comprehension
Andreas Schramm & Michael C. Mensink

Chapter 10
Statistical sequence and parsing models for descriptive linguistics and psycholinguistics
Gerold Schneider & Gintaré Grigonytė

Name index

Subject index
Introduction

Sarah Chevalier, Anne-Christine Gardner, Alpo Honkapohja, Marianne Hundt, Gerold Schneider & Olga Timofeeva
University of Zurich

The selection of articles presented in this volume springs from the Third Conference of the International Society for the Linguistics of English that took place in Zurich, Switzerland, on 24–27 August 2014, and in particular from the special theme of the conference “Building Bridges – Inter- and Intradisciplinary Research.” Thus, on the one hand this book aims at providing a cross-section of current developments in English linguistics, and on the other it does so by tracing recent approaches to corpus linguistics and statistical methodology, by introducing new inter- and multidisciplinary refinements to empirical methodology, and by documenting the on-going emphasis shift within the discipline of English linguistics from the study of dominant language varieties to that of post-colonial, minority, non-standardised, learner and L2 varieties. Among the key focus areas that define research in our field today, we have limited our selection to four: corpus linguistics, English as a global language, cognitive linguistics, and second language acquisition. Most of the articles in this volume concentrate on at least two of these areas and at the same time bring in their own suggestions towards building bridges within and across sub-disciplines of linguistics and beyond. In the introductory sections that follow, we summarise contemporary advances in the four chosen topics and highlight the links between the articles both within and across the topics. The summaries of the individual papers are provided in the second part of the introduction.

1. Corpus linguistics

All the papers in our volume bring in methodological innovation to enlighten our understanding of linguistic variation and change, by using corpus resources in creative ways or by introducing/compiling new corpora, by applying statistical methods, or by borrowing new methodologies, e.g., from cognitive linguistics or second language acquisition.
Corpus linguistics has constantly widened and expanded its field of application, as Sampson and McCarthy note:

Fifty years ago, corpus linguistics was an obscure and highly specialized minority activity. Since then, slowly at first but in the last ten years almost explosively, it has widened out to provide virtually every approach to the study of language, humanistic or technical, with new methods and new insights. By now, many agree with a widely quoted remark by Hoey in 1998: “Corpus Linguistics is not a branch of linguistics, but the route into linguistics.” (Sampson & McCarthy 2004: 4)

Since these lines were written, corpus linguistics has expanded further. The use of corpora has revolutionized many branches of linguistics, not least by means of bringing statistics, empirical methods and machine learning to the discipline. This has occurred most notably in computational linguistics, but also in psycholinguistics (Janda 2013), and possibly in linguistics as a whole, as the crucial role of frequency in grammaticalization and language processing has been recognized (Bybee 2007: 337).

At the same time, perceived gaps between the sub-disciplines are closing and bridges are being built, such as the one, for example, between corpus and computational linguistics. Anticipated by Tognini-Bonelli’s (2001) description of corpus- or data-driven approaches, statistical modelling approaches are now reaching centre-stage in corpus linguistics and computational linguistics alike. Natural language processing and computational linguistics started off as engineering approaches, often with simplistic linguistic assumptions, and with a completely task-oriented perspective. But approaches have matured, computing power increased, and the size and range of corpora which are available today have made it possible to test ever more complex models and linguistic theories (for example, with a complex statistical model for which only a computer can keep track of all the factors). Further, they have made new areas of research possible, for example Learner Corpus Research, to which the contributions by Laitinen, by Wärnsby et al., by G. Schneider & Grigonytė and by Gilquin are dedicated.

In complex models which recognize the importance of frequency, it is uncontroversial that statistics are becoming the cornerstone. But Gries (2012) laments that university curricula in linguistics do not give enough attention to teaching statistical methods (Gries 2010: 123). Gries (2010, 2015) and others (e.g., Evert 2006) recommend the use of multivariate regression, mixed effect models and other advanced statistical approaches, and observe that they remain underused. This deficit is addressed in the current volume in the contributions by A. Schneider and by Krug, Schützler and Werner, who use mixed-effect regression (the latter also a phylogenetic data-driven clustering approach). A data-driven approach is also taken by Bruckmaier, who uses what she calls “a radical corpus-driven approach”
(122), in which separate word forms are used as the basic unit of analysis rather than grouped together under a lemma, and the categories arise from the data.

The contribution by G. Schneider and Grigonytė, on the other hand, uses a computational linguistics tool which heavily employs statistics, namely an automatic syntactic parser as tool and as model. The use of parsers for large-scale analysis with sufficient accuracy has become possible recently, as van Noord and Bouma (2009) point out.

2. Global English

Recent advances in corpus linguistics and the use of statistics are also highly visible in the second focus area of the volume: variation and change in the language of people who regularly use English, whether as a first, second or foreign language (the latter groups greatly outnumbering the former) (see Crystal 2008). Since the 1980s, variation and change in English on a global scale has been the focus of many publications, e.g., in the Varieties of English Around the World series or in specialized journals such as World Englishes (typically with a more applied focus, including aspects of teaching and language politics) or English World-Wide (with a clear focus on sociolinguistics and the description of world Englishes). Over the years, the field has become an established part of English socio- and variationist linguistics, moving from the description of fairly well-known varieties to lesser-known varieties, e.g., English in Tristan da Cunha (Schreier 2003). Various scholars have developed theoretical frameworks for the study of world Englishes (such as Strevens’ 1980 “world map of English” or Kachru’s 1992 “three circles model”). More recently, Edgar Schneider (2003, 2007) proposed a dynamic model for the evolution of postcolonial Englishes that has made a significant impact in the field, as evidenced by the many publications it has spawned (see, e.g., the contributions in Buschfeld et al. 2014). Another milestone for the development of the field was the launching of the International Corpus of English (ICE) project in the 1990s: ICE has enabled detailed, corpus-based descriptions of different Englishes, increasingly also with a cross-varietal perspective (see, e.g., Hundt & Gut 2012).

Like other areas in English linguistics, the study of world Englishes has lately seen what we might want to call a “numerical” turn, i.e., variation is modelled in probabilistic terms (typically based on substantial amounts of corpus evidence). While for the longest time, investigation of second-language varieties of English and learner English was conducted in two separate fields, using separate frameworks and methodologies, scholars have started to bridge what has been called a “paradigm gap” and begun to work towards a rapprochement of World Englishes and SLA research.
The globality of English invites our readers to reconsider time and place parameters of language change on at least three different levels: a) inner circle: the idiosyncrasies of English among native English communities in a foreign country are addressed by Hirano & Britain; b) World Englishes: Ghanian, Jamaican, Maltese, Puerto Rican, Singaporean and more are tackled by Krug, Schützler and Werner, Green, Bruckmaier, A. Schneider; c) English as a second or foreign language is used as data by Gilquin, Laitinen, Wärnsby et al., Schramm and Mensink, and G. Schneider and Grigonyté.

3. Cognitive linguistics

One of the most important bridges built by this volume is from corpus linguistics to cognitive and psycholinguistics. Many of the articles are also linked by the third umbrella topic, the underlying topic of cognitive and psycholinguistic explanation, which has always been the holy grail and ultimate justification of linguistic research, all the way from Antiquity (Plato’s concern was whether language was a man-made artefact, or supernatural in origin), German philosophy (Wittgenstein’s *Sprachspiel* stresses the social aspects of cognition and that meaning presupposes cognition and use) via Chomsky’s I-language as a biologically based feature of the brain (Chomsky 1986) up to current efforts of using computational models for language acquisition and processing (e.g., Lenci et al. 2014). It is increasingly being recognized that the scientific study of language needs a common effort by linguists, neuro-scientists and psychologists (Walenski & Ullman 2005: 328). Cognitive linguistics considerations unite contributions by Schramm and Mensink and G. Schneider and Grigonyté; corpus compilation is the focus in Laitinen and Wärnsby et al.

4. Second language acquisition

Learner language (L2) offers an exciting bridging position between cognitive linguistics and corpus linguistics. L2 speakers are challenged precisely because they have not been exposed to as much data as native speakers, as Pawley and Syder (1983) point out. As such, they do not have the opportunity to acquire the same stock of semi-formulaic utterances as native speakers do.

The stock of lexicalized sentence stems known to the ordinary mature speaker of English amounts to hundreds of thousands. In addition there are many semi-lexicalized sequences, for just as there is a continuum between fully productive
rules of sentence formation and rules of low productivity, so there is a cline between fully lexicalized formations on the one hand and nonce forms on the other. (Pawley & Syder 1983: 192)

The close correlations between frequency and expectations, for example semantic expectations associated with words, which have, e.g., been shown by Schulte in Walde and Melinger (2008), is another connection between psycholinguistics and corpus linguistics which has not been investigated enough to date. But the connections between the fields are increasingly being recognized. Gries (2012: 47) states that “cognitive approaches to language are not only compatible with much recent work in corpus linguistics, but also provide a framework into which corpus-linguistic results can be integrated elegantly.” The current volume aims to deliver a contribution to this research trend.

Five papers in the present volume are concerned with L2 English, four of which rely on corpus linguistic methodology. These contributions push the boundaries of traditional distinctions and methods, present new resources, and shed light on areas which until now have remained underexplored. Gilquin, for example, builds a bridge to global English (e.g., Hundt & Mukherjee 2011) in her comparison of EFL (English as a foreign language) with ESL (English as a second language). Further, she disregards the usual connection of ESL varieties to Kachru’s (1992) outer circle and EFL varieties to his expanding circle, by stating that Sweden is “widely recognised” as an ESL country. Laitinen, as well as Wärnsby et al. present new corpora that fill gaps in the current range of L2 English corpora available. Laitinen presents two corpora of advanced non-native English texts in a wide variety of genres, a resource hitherto lacking. Wärnsby et al. likewise introduce a new corpus resource, consisting of multiple drafts of student texts and feedback on these. The latter resource will enable researchers in linguistics, pedagogy and writing alike to gain a greater understanding of the process of acquiring writing skills in L2 English. Schramm and Mensink, by contrast, focus on comprehension rather than production and cast light on a little-examined area in learner English, namely learners’ understanding of grammatical aspect in narratives. Finally, G. Schneider and Grigonytė show that learners (even advanced ones) have limited command of formulaic language, and their production in turn is more difficult to process both for native speakers and for automatic parsers. This suggests that the latter can be used as a model of human language processing, or at least provide clues about how humans do this, and why they make comprehension mistakes.
5. The individual contributions

The first two contributions offer a global perspective on synonymous lexical items (e.g. HAVE GOT TO vs. HAVE TO, or anticlockwise vs. counterclockwise) which are traditionally associated with British and American English, respectively. Both contributions provide a statistical analysis of lexical usages in selected varieties of English, yet each reveals a different research focus: while Hirano and Britain study the role of accommodation and social networks in changes concerning traditional preference patterns observable in a group of English speakers in Japan, Krug, Schützler and Werner examine the relative distance between four varieties of English in terms of similarities and differences in lexical usage and uncover a move towards globalisation in second-language varieties. In “Accommodation, dialect contact and grammatical variation: Verbs of obligation in the Anglophone community in Japan”, Hirano and Britain apply a social network approach to explore the intermediate stage in dialect contact, i.e., “the stage between fleeting accommodation and permanent linguistic change” (14). Focusing on changing usages of the verbs of obligation MUST, HAVE TO, HAVE GOT TO and GOT TO, of which HAVE GOT TO is predominantly associated with British English (BrE) and HAVE TO with North American Englishes, they examine on the basis of spoken data whether young speakers of particular varieties of English (BrE, AmE and New Zealand English) converge towards each other after prolonged contact with other English varieties in Japan where they are based as teachers or during their university studies. Hirano and Britain analyse the lexical choices of individuals at the beginning of the contact period and again after one year, and detect, for instance, a pattern of divergence between AmE and BrE speakers, the latter increasing their usage of HAVE GOT TO; however, when these BrE speakers have strong ties with Americans this increase is less pronounced. In a final step the authors consider grammatical contexts to account for the lack of convergence observed in their data.

In their contribution “Patterns of linguistic globalization: Integrating typological profiles and questionnaire data”, Krug, Schützler and Werner investigate the differing usage of 68 lexical binaries such as tap and faucet in four varieties of English: BrE and AmE, as well as Maltese English (MaltE) and Puerto Rican English (PRE). By combining questionnaire-elicited acceptability ratings for these lexical items with aggregative analysis and visualisation through phenograms, and by using regression models, the authors provide new insights into the interrelatedness between, as well as lexical variation within, these varieties. They identify two clusters of lexical usage, with BrE and AmE at opposite poles, and MaltE and PRE situated closely to that variety by which they have been influenced historically and politically, i.e., BrE and AmE respectively. Furthermore, they observe that the second-language varieties are closer to each other in terms of lexical preferences,
and feature considerably more internal variation, than BrE and AmE. There is also evidence for apparent-time changes in MaltE and PRE with a move towards free variation and away from exonormative orientation, which the authors argue to be globalising tendencies – a process also observable in the first-language varieties.

Exploring a larger number of varieties of English than formed the basis of the first two contributions, Green concentrates on the global distribution of two lexical items and studies “The substitutability and diffusion of want to and wanna in World Englishes” by drawing on data from the Corpus of Global Web-Based English (GloWbE). He examines phonological, grammatical, semantic and pragmatic distribution patterns (including, but not restricted to, occurrence in clause-final positions; co-occurrence with specific modals or subordinating conjunctions; use for senses like “intention”, “obligation” and “probability”; expression of speaker projection and advisory use) in order to ascertain in which contexts WANT TO and WANNA are used, and whether they are affected by any particular contextual constraints. In addition, Green provides a survey of the diffusion of these patterns in the twenty varieties of English represented in the corpus. He finds for instance that WANT TO and WANNA are indeed substitutable in many contexts, but that WANNA rarely occurs clause-finally in the middle of a sentence, and that WANT TO is preferred to express a sense of “obligation”. Considering that WANT TO is attested in a wider range of usages in more varieties of English than WANNA, Green concludes that the diffusion of WANNA, which appears commonly in informal contexts, is still an ongoing process.

In her paper “Dialect contact influences on the use of get and the get-passive”, Elisabeth Bruckmaier studies different forms of the highly frequent and often stigmatized verb GET, focusing particularly on its occurrence in passive constructions in Singaporean and Jamaican English. The methodological contribution made by the article is the use of a bottom-up corpus driven approach, in which the individual word forms are used as the basic unit of analysis, whereas in previous studies the forms have often been conflated under the lemma or left unspecified. Bruckmaier, in contrast, considers forms such as get, gotten, gotta as separate lexical items, and also takes into account two passive forms particular to Singaporean English the kena-passive and give-passive. For data, she mainly uses three ICE-corpora, ICE-Great Britain (ICE-GB), ICE-Jamaica (ICE-JA), and ICE-Singapore, contrasting the results occasionally with other ICE corpora as well as the Corpus of Global Web-based English (GloWbE). The results display Jamaica in the sphere of American English and Singapore in the sphere of British English.

Agnes Schneider investigates variation in marking future time in Ghanaian English, including the modal will and its variants will, ’ll, won’t and BE GOING TO and the variants be going to and be gonna. The corpus consists of 144,000 words of spoken conversations by Ghanaian speakers with a minimum of secondary school
education, recorded by the author during her fieldtrips to Ghana in 2002, 2008 and 2010. These are contrasted with data consisting of 180,000 words of spoken conversations of British English from the ICE corpus. The methodology consists of a multivariate mixed effect logistic regression model combined with discourse analysis. The results reveal *be going to* as considerably less frequent in Ghanaian English than British English, and that a number of constraints which affect its use, for example in British English *be going to* is almost as frequent as *will* with 2nd and 3rd person subjects, whereas in Ghanaian English *will* is preferred with all subjects. A. Schneider suggests a number of reasons for this, including (incorrect) association of *be going to* as a feature of American English, the influence of semantic systems of indigenous Ghanaian languages and most importantly for the present collection: a bridge to SLA, since the actual constraints that influence the choice of future time expressions by native speakers are not taught in a classroom setting.

In “Ongoing changes in English modals: On the developments in ELF” Laitinen explores new ways of investigating English in the expanding circle. Drawing on methods from historical linguistics and variationist sociolinguistics, he investigates how on-going grammatical changes are adopted in advanced non-native English. To this end, Laitinen and his colleagues are compiling two new corpora of texts written by advanced non-native writers of English in Sweden and Finland. Unlike existing non-native corpora, these corpora cover a wide variety of written genres, such as administrative emails, tweets, and professional blogs. Such a corpus design is timely since much English used in non-native contexts today is found in the form of online texts. Based on a comparison between data in these corpora and those in existing corpora of native English, Laitinen reports on a case study concerned with recent and ongoing changes in core modals (e.g., *can*) and emergent modals (e.g., *have to*). His research shows that while advanced non-native English follows the American trend of a decrease in the usage of core modals, there is considerable polarisation in non-native usage, i.e., many of the items which have undergone substantial recent increases or decreases in native use exhibit more extreme patterns in non-native use.

Wärnby, Kauppinen, Eriksson, Wiktorsson, Bick and Olsson likewise present a new corpus of texts produced by L2 writers of English. The Malmö University-Chalmers Corpus of Academic Writing as a Process (MUCH) is being compiled in order to analyse the writing process and the influence of feedback on it. A particularly innovative aspect of this corpus lies in its inclusion of multiple drafts of texts. The project aims to include a large number of student papers of different performance levels in different drafts, as well as student metadata, and peer and instructor feedback. In this way, it will be possible to study how texts develop and change during the writing process and how feedback has an impact on the process,
which may result in pedagogical development of writing in higher education. The compilers plan to annotate the papers and feedback in a way that makes possible complex queries, and explain that they will create a search interface tailored to the needs of the writing research community. The latter point is of considerable significance since MUCH will be an open corpus available to, and designed to help collaboration between, scholars in writing, pedagogy and language alike.

Remaining within the field of advanced learner English, Gilquin’s contribution examines “Discourse markers in L2 English: From classroom to naturalistic input”. Gilquin hypothesises that as discourse markers are not widely taught, EFL learners will have problems with this feature of English, since their main exposure to the target language is in the context of the classroom. Gilquin suggests that even advanced learners will underuse (or misuse) discourse markers. Her hypothesis is tested via a comparison of the Louvain International Database of Spoken English Interlanguage (LINDSEI), a corpus of speech produced by advanced learners of English, with a similarly designed corpus of native (British) English, the Louvain Corpus of Native English Conversation (LOCNEC). The study reveals a general underuse of discourse markers (except for well, which is overused) and certain instances of misuse. By contrast, exposure to more naturalistic input, measured by stays in English-speaking countries, helps L2 speakers use discourse markers more appropriately. Gilquin also compares foreign versus institutionalised second-language varieties of English, finding that ESL speakers tend to better approximate native speaker usage of discourse markers in line with their higher exposure to naturalistic English.

The next chapter co-authored by Andreas Schramm and Michael C. Mensink opens the last set of two papers that bring together language acquisition and language processing. Theirs is a contrastive investigation of perfective simple past vs. imperfective past progressive and the role that aspect plays in narrative comprehension and in second language development. The aim is to compare native and non-native speakers of English in their processing of aspectual meanings in narrative texts. Do advanced learners process such meanings, are they able to do so without explicit instruction, and how much their performance differs from native-speaker processing of aspect? A finely designed experiment tests 25 L1 Arabic advanced learners of English and 32 native speakers of English in their comprehension of grammatical aspect and the effects it has on their working and long-term memory. The analysis of the results reveals that non-native readers, even at the advanced stage, differ greatly from native speakers, in that they do not seem to notice aspectual meanings and fail to re-instate them both during and after reading. Thus it appears that the aspectual input is not registered cognitively, and therefore the study suggests that implicit learning of aspectual meanings should seem unlikely. This crucial divide between native and non-native speakers comes
up again in the final chapter in which human comprehension is compared to computer parser confidence scores.

Gerold Schneider and Gintaré Grigonytė’s is in many respects one of the most daring contributions in the whole collection, and also one that offers both a critical summary of statistical methods employed in linguistics today and a likely scenario of how and where to our field will be developing in the next years. They start out their survey with significance tests and three assumptions that are typically taken for granted by linguists: the assumption of random, “normal” distribution of data, the assumption of data independence from other factors, e.g., region from genre or gender from social background, and the assumption of speakers’ free choice. All the three can be shown to have limited validity. Consequently the use of multifactorial predictive models (such as regression models) that can predict significant variables and the alternations between them is advocated as a highly reliable tool. The problem of free choice, or rather the lack thereof, at all levels of language, however, remains. Empirical research shows that formulaic language, expected continuations, and chunking always have processing advantages for native speakers. Language learners, however, acquire formulaic language relatively late. Their subtle failures, then, must produce unexpected features that increase processing load and ambiguity of utterances. This hypothesis is tested with three global language-processing models: surprisal, a surface-based word-sequence model, POS tagging, a pre-terminal surface model, and a syntactic parser, a hierarchical model, and eventually confirmed – model fits and parser scores are lower for L2 utterances. This is in line with previous studies on native-speaker comprehension of learner non-formulaic features. This leads Schneider and Grigonytė to a conclusion that automatic and human parsers work similarly and that, ultimately, a type of syntactic parser is the candidate for a psycholinguistics model.

Acknowledgements

The editors are very grateful to the contributors of this volume and to the co-authors of the introduction for their willingness to cooperate and share their work and ideas, and their prompt responses to questions and deadlines. We must also thank the general editors of the SLCS series and the team at Benjamins for their support in the production of this book. Our special thanks go to our editing assistant Anna Fuhrer.
References


CHAPTER 1

Accommodation, dialect contact and grammatical variation
Verbs of obligation in the Anglophone community in Japan

Keiko Hirano and David Britain
University of Kitakyushu / University of Bern

The present study investigates dialect contact and linguistic accommodation in the use of verbs expressing obligation (such as MUST, HAVE GOT TO, HAVE TO and GOT TO) among native speakers of English resident in Japan, using a social network approach. Approximately 500 tokens were extracted from conversations between 39 native speakers of English from England, the US and New Zealand, recorded in single-nationality dyads, both immediately upon arrival in Japan and after a period of one year. Statistical analysis revealed that the informants from England actually diverged from the forms typically used by the Americans. The results, however, demonstrate the importance of social network strength in accounting for the consequences of dialect contact and short to medium-term linguistic accommodation.

Keywords: dialect contact, linguistic accommodation, social network approach, verbs of obligation

1. Introduction: Accommodation and contact

Dialect contact approaches to linguistic change investigate the consequences for linguistic structure of interaction between speakers of distinct but nevertheless mutually intelligible language varieties (Britain in press; Trudgill 1986). A good deal of the foundational research in this domain has examined either the linguistic ramifications of the fleeting contact that takes place, for example, in service encounters or other short-term speaker interactions (e.g. Coupland 1984), or the outcomes of contact that results from acts of long distance and/or large scale migration or other significant acts of mobility by speakers of often quite radically different varieties (e.g. Trudgill 2004).
On the one hand, therefore, research, drawing largely upon empirical investigations of short-term contact, has argued that there is a “general and seemingly universal (and therefore presumably innate) human tendency to ‘behavioural coordination’” (Trudgill 2004: 27–28) leading to interactional convergence, and, on the other hand, drawing on work particularly on the formation of post-colonial language varieties, especially Englishes, that there is a typologically coherent set of linguistic changes which occur, over a couple of generations, as a result of the coming together of distinct dialects following settler migration and other larger-scale mobilities. To explain the latter, dialect contact approaches rely on the routinisation of the kinds of convergent changes that are characteristic of the former. Researchers argue, therefore, that temporary linguistic convergence typically occurs when people interact, even briefly, but that if such interactional synchronisation to speakers of distinct varieties becomes regular and routine, convergent linguistic features may be regularly adopted as variants in people’s repertoires, or even become over time, and in the right conditions, permanent features of their dialect (Trudgill 1986).

Whilst much of the work in the (fleeting short-term) speech accommodation paradigm is experimental and readily replicable, and a good deal of the theorising about the outcomes of long term dialect contact has come from post-hoc investigations and descriptions of newly formed postcolonial dialects (Schneider 2007; Trudgill 2004), it has been much more methodologically challenging (see Meyerhoff 1998) to investigate the crucial intermediate stage – the stage between fleeting accommodation and permanent linguistic change. What happens after a year of accommodation to dialects that are distinct to your own? After two years? After ten? As the longitudinal depth of such investigations becomes greater, so too do the empirical challenges, because the possibility of tracking groups of migrant individuals over ever longer periods of time becomes impractical. Researchers have attempted to tackle the investigation of this intermediate period through studies of second dialect acquisition – studies which look at the linguistic ‘success’ of individuals who, as a result of migration, are faced with acquiring an ambient target dialect different to their native variety (e.g. Chambers 1992). While we have learnt a great deal about the sorts of linguistic features which are readily and not so readily adopted in such circumstances, these studies tend to be based, understandably, on very small samples and/or on members of the same family, and often only post-migration recordings of the speakers are available. Engineering appropriate, neat, ‘scientific’ samples (where everyone moved from the same community to the same community at the same time) of recordings gathered both before departure and at regular intervals thereafter, is the (not entirely successfully accomplished) stuff of extremely well-funded and durable TV series such as *7-Up*, rather than routinely achievable by sociolinguistic fieldworkers. So while a number of studies
show accommodation in the short term and a typology of outcomes of contact in the long term, methodological difficulties have prevented the development of a large and robust literature that examines this crucial intermediate stage – the linguistic outcomes of contact that is, in a sense, medium term and which might provide evidence of the ‘becoming-permanent’ of dialect convergence.

Long term dialect contact studies also take seriously the potential role of social network integration in encouraging (or not) both linguistic change and the development of new community norms, including linguistic norms, in the post-contact sociolinguistic aftermath, following the principle, from James Milroy and Lesley Milroy, that strong local social network ties act as norm enforcement mechanisms (e.g. Milroy & Milroy 1985). The theoretical influence of the network model has been considerable, but this has not actually been matched by an equally considerable number of studies attempting to empirically capture the influence of networks on language data (for a review of empirical work in this area see Milroy & Llamas 2013). In dialect contact research especially, social networks have largely served theoretical rather than empirical functions (though see Bortoni-Ricardo 1985 for an interesting and relatively early exception).

2. Aims of the study

Here we attempt to add to evidence on the intermediate stage, by reporting on a medium-term real-time panel-study (see Sankoff 2005) of dialect contact and grammatical variation, where spoken data were collected from almost 40 speakers from different dialect backgrounds both at the start of a period of contact, and then a year later, and in which social network measurements were designed to enable correlations between different strengths of community integration and linguistic usage. The community in question consists of young English speakers of different nationalities in Japan who spent at least a year working there as teachers. Do these speakers (English, American and New Zealand, in our broader study) demonstrate convergence towards each other after a year in Japan, and if so, is this convergence shaped by social network ties with other speakers in the community, both native non-native?

In earlier research on phonological variables in the dataset, Hirano (2013) demonstrated that indeed there was convergence in cases where people had strong network ties with people from other Anglophone countries. For example, American speakers with strong social network ties to British people were more likely to use glottal /t/ intervocally and English informants with strong network ties with Australians and New Zealanders were less likely to use glottal stops intervocally. The study also showed that if people tended to stay within their
“nation group”, this hindered the adoption of forms from other varieties. For example, those English informants with strong social network ties to British people were less likely to flap their intervocalic /t/ and more likely to glottalise their intervocalic /t/ than those who did not have such strong networks with other Brits.

Here we ask whether we can observe similar subtle convergences, after just a year, when examining grammatical variables. In order to do this, we need to examine linguistic variables for which, among the degree-educated speakers in our corpus, there are clear differences between, for example, American and British English.

For the present analysis, therefore, we have chosen to examine verbs of obligation: MUST v HAVE TO v HAVE GOT TO v GOT TO, as shown in examples (1) to (4) respectively.

1. we must do it like every … three or four months

2. you have to have a steady hand

3. I’ve got to go to school

4. you got to start from somewhere

HAVE TO is strongly associated with North American Englishes, HAVE GOT TO with British English, with Australasian Englishes using HAVE GOT TO more than American varieties, but less than British ones (e.g. Tagliamonte 2013: 137; Collins 2005: 261). In conducting this analysis, we also take into consideration one linguistic constraint that has often been shown to operate on variation across this set of verbs: type of obligation (subjective versus objective) (see Section 4.2 for more details). Before moving to the analysis, however, we briefly introduce the Anglophone community in Japan that is the subject of our study.

3. The Anglophone community in Japan

The speech community under investigation here consists of native speakers of English from English-speaking countries. They are living temporarily in Japan as language instructors working for state schools or private institutions. The start of English education in Japan dates back to the latter half of the 19th century, with English becoming a compulsory subject in the vast majority of secondary schools after World War II (Honna 2009: 119). Today, most secondary schools and universities teach English as the first foreign language (Okuno 2007: 38).

There are around 100,000 native speakers of English either teaching or studying in Japan (National Statistics Center 2015). The majority of those who come to Japan as teachers of English at state schools are participants on the Japan Exchange
and Teaching [JET] Programme,\(^1\) which is sponsored by Japanese ministries\(^2\) with the purpose of increasing mutual understanding between the people of Japan and the people of other nations. The programme employs young university graduates from overseas on fixed term contracts. Since the beginning of the JET Programme in 1987, there have been more than 60,000 participants from 63 countries around the world. The number of JET participants per year reached over 6,000 in 2000, and the most recent report indicates that over 4,000 people from 42 countries are participating in the programme in 2014–2015 (CLAIR 2014). According to the eligibility criteria published on the JET official homepage (CLAIR 2014), applicants for assistant language teachers whose primary task is to assist Japanese teachers of foreign languages in classrooms at schools must hold a bachelor’s degree, must “be adept in contemporary standard pronunciation, rhythm and intonation”, and “possess excellent language ability that can be applied accurately and appropriately”.

The participants typically stay in Japan for one to three years, and they are constantly replaced by new arrivals. Their relationships, therefore, are often established on a short-term basis, but they are linked with people from a wide range of social contexts. According to the information we collected from our informants, almost all tend to develop social network ties with other English teachers like themselves. Social interaction, therefore, often takes place in an intra-group context. In this context, they mix with native speakers of other regional varieties of English. There are also many opportunities for interacting with Japanese people outside of work. Some may take lessons in various aspects of Japanese art, culture and martial arts while others may join a local sports team. The JET participants are also expected to engage in local international exchange activities. They would use both English and Japanese according to the English proficiency of the people they interact with and their own Japanese ability. It is this group that we have used in our study to examine medium-term convergence.

---


2. The JET Programme is sponsored by the Ministry of Internal Affairs and Communications, the Ministry of Foreign Affairs, the Ministry of Education, Culture, Sports, Science and Technology, and the Council of Local Authorities for International Relations (CLAIR 2014).
4. Methodology

4.1 Informants and data

Thirty-nine informants of three nationalities participated in our study: 15 English, 11 Americans and 13 New Zealand nationals. All had received higher education and were working in Japan as language teachers. They included 36 assistant language teachers on the JET Programme and three English conversation instructors employed by private language schools. They were aged between 21 and 34 years old at the time of the first data collection, and 24 years old on average. In terms of social background – age, educational achievement, occupation and income in Japan, for example – the informants represent a relatively homogeneous group.

The data used for this study were collected in three different prefectures of Kyushu – Fukuoka, Saga and Kumamoto – as shown in Maps 1 and 2, from the same informants on two separate occasions. The first dataset was collected immediately after the informants’ arrival in Japan and the second set was collected about a year later. Some of them returned home after a year but most of them remained in Japan for a second year. This longitudinal study was conducted to compare and contrast the results of the two datasets and to trace the course of changes in real time. In both sessions, natural, spontaneous conversations between two native speakers of English from the same country were recorded in their homes in a relaxed atmosphere for 45 minutes. In most cases the two people in each pair were friends. A total of 34 hours of speech were used for the present study.

The current research used a method designed to elicit naturally occurring conversation from the informants. The interviewer was not present while the informants were being recorded in order to lessen the possibility of speech modification that might result from the presence of a researcher from Japan who is a non-native speaker of English. In order to minimise any possibility of short-term accommodation – the risk of linguistic changes being caused by the conversation partner – the informants were paired with someone from the same country, not someone who spoke a different variety of English. Any linguistic changes observed between the first and the second datasets, therefore, were less likely caused by the conversation partner but more likely caused by the linguistic environment in which informants had daily conversations with various native speakers of English or non-native speakers of English throughout the preceding year in Japan.

3. There is one exception. One NZ informant had an Australian partner whose data is not used in this study.
Map 1. Map of Japan

Map 2. Map of Kyushu
4.2 Tokens and analysis

We extracted all tokens of MUST, HAVE TO, HAVE GOT TO and GOT TO from our two datasets. As shown in Table 1, we collected 227 tokens from the first dataset and 254 tokens from the second dataset, a total of 481.

<table>
<thead>
<tr>
<th>Data</th>
<th>English</th>
<th>Americans</th>
<th>New Zealanders</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st dataset</td>
<td>82</td>
<td>63</td>
<td>82</td>
<td>227</td>
</tr>
<tr>
<td>2nd dataset</td>
<td>116</td>
<td>60</td>
<td>78</td>
<td>254</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>123</td>
<td>160</td>
<td>481</td>
</tr>
</tbody>
</table>

One of the linguistic constraints, found to especially shape variation in verbs of obligation, was taken into consideration for each token: Type of obligation. Type of obligation considers the source of the obligation, who is doing the obliging, the speaker or someone else.

(5) I don’t see why I have to get an x-ray

(6) I must admit I haven’t watched the TV much

The authority of objective obligation, as in (5), comes from “a source external to the speaker”, while the authority of subjective obligation, as in (6), comes from “the speaker him-/herself” (Tagliamonte & D’Arcy 2007: 65). In (5), the obligation to have an x-ray comes from a medic, not the speaker herself. In (6), however, it is the speaker who is imposing on herself the requirement to admit something. It has generally been found that levels of HAVE TO are higher for objective than for subjective obligation, with the reverse the case for HAVE GOT TO (e.g. Tagliamonte 2013: 142–145; Tagliamonte & D’Arcy 2007: 76–77).

To provide results of the linguistic data analysis in relation to social network, SPSS version 22 was used and statistical tests including Pearson correlation and multiple regression were performed. To perform these analyses with SPSS, it is required that each person provides sets of scores. The percentage use of a particular variant for each informant, therefore, was used in the statistical analyses instead of the raw frequencies for each variant.

4. Tokens found in negative and interrogative sentences, in which some variants either occur very rarely or not at all, were not included in the present analysis.
4.3 Social networks

This paper attempts to demonstrate that the social networks of informants play an influential role in their convergent linguistic behaviour. Milroy (1980) took density\(^5\) and multiplexity\(^6\) into account to measure the strength of social networks, but native speakers of English who have recently joined a fluid, mixed dialect contact situation cannot be expected to have long-standing, dense, multiplex, close-knit networks within the new context since most of them stay temporarily and are constantly replaced by new arrivals. The present study took into account speakers’ self-assessed closeness to other Anglophones or with Japanese, the frequency of meetings and telephone conversations with them, and network size. It then developed a number of quantitative indices that would enable us to correlate linguistic shifts between the two datasets and network strength.

The informants had a short interview with the researcher at the end of the second data collection session and were asked about people with whom the informants had a close relationship and regular contact in their daily life in Japan. A set of 11 questions, shown in Table 2, were used to elicit information about the informants’ social networks. A score for each relationship was calculated using the rank order of closeness and the frequency of meetings and telephone calls with the person as follows:

\[
\text{Score for each relationship} = \text{rank order score} \times (\text{score for meeting frequency} + \text{score for telephone call frequency})
\]

These individual relationship scores were then grouped into different social network categories.

A network index score\(^7\) represents the strength of the individual social network of the informant.\(^8\) First of all, the social network of each informant was grouped into two sub-networks: a *network with native speakers of English* and a *network with non-native speakers of English*. The network with native speakers of English was further divided into three sub-networks: *British*, *North American* and *Australasian networks*. Additionally, a network index score with *English teachers* including native speaker English teachers and Japanese teachers of English was

---

5. The density of a social network depends on the degree to which the members who form the social network know each other.

6. The multiplexity of a social network depends on the extent to which individuals are linked to one another by more than one relationship category.


8. Note that ‘a network with a high network index score’ in the current study does not mean a close-knit network, a term used by Milroy (1980) to describe a dense and multiplex network.
also created. The index scores of these networks are used to examine any relationship with informants’ choice of variants of verbs of obligation. So, for example, we could examine, in the case of the New Zealand group, the linguistic effect of the strength of their network ties with North Americans, British or other New Zealanders and Australians, and, additionally, with Japanese. The informants’ average index scores of sub-networks for both native speakers of English and non-native speakers of English are shown in Table 3.

Table 2. Questions about informants’ social networks

<table>
<thead>
<tr>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. names or initials of close friends</td>
</tr>
<tr>
<td>Q2. sex of each person</td>
</tr>
<tr>
<td>Q3. frequency of meetings with each person</td>
</tr>
<tr>
<td>Q4. frequency of telephone calls with each person</td>
</tr>
<tr>
<td>Q5. relationship with each person (friend, colleague or neighbour)</td>
</tr>
<tr>
<td>Q6. whether or not each person is an English teacher in Japan</td>
</tr>
<tr>
<td>Q7. nationality of each person</td>
</tr>
<tr>
<td>Q8. years spent in Japan by each person (for non-Japanese only)</td>
</tr>
<tr>
<td>Q9. main language used with each person</td>
</tr>
<tr>
<td>Q10. interconnecting close ties with one another</td>
</tr>
<tr>
<td>Q11. ranking of these relationships</td>
</tr>
</tbody>
</table>
Table 3. Average index scores of sub-networks according to informant nationality

<table>
<thead>
<tr>
<th>Social networks</th>
<th>English</th>
<th>American</th>
<th>NZ</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Networks with native speakers of English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) British</td>
<td>34.7</td>
<td>8.7</td>
<td>18.5</td>
<td>20.6</td>
</tr>
<tr>
<td>(2) North American</td>
<td>25.2</td>
<td>38.4</td>
<td>31.8</td>
<td>31.8</td>
</tr>
<tr>
<td>(3) Australasian</td>
<td>10.7</td>
<td>2.3</td>
<td>19.1</td>
<td>10.7</td>
</tr>
<tr>
<td>(4) Other NSsE</td>
<td>1.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(5) English teachers</td>
<td>(70.6)</td>
<td>(44.3)</td>
<td>(66.5)</td>
<td>(60.5)</td>
</tr>
<tr>
<td>Subtotal</td>
<td>71.8</td>
<td>49.4</td>
<td>69.4</td>
<td>63.5</td>
</tr>
<tr>
<td>Networks with non-native speakers of English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6) Japanese</td>
<td>29.6</td>
<td>54.5</td>
<td>29.9</td>
<td>38.0</td>
</tr>
<tr>
<td>(7) Other non-NSsE</td>
<td>1.5</td>
<td>0.0</td>
<td>2.3</td>
<td>1.3</td>
</tr>
<tr>
<td>(8) English teachers</td>
<td>(10.5)</td>
<td>(22.0)</td>
<td>(5.7)</td>
<td>(12.7)</td>
</tr>
<tr>
<td>Subtotal</td>
<td>31.1</td>
<td>54.5</td>
<td>32.3</td>
<td>39.3</td>
</tr>
<tr>
<td>English teachers (5) + (8)</td>
<td>81.1</td>
<td>66.3</td>
<td>101.7</td>
<td>80.0</td>
</tr>
</tbody>
</table>

5. Results

The results section will first look at the overall distribution patterns in the two datasets. Second, the effect of obligation type on the distribution of variants will be examined, and then the impact of social network on variation will be explored.

5.1 Overall distribution of variants of verbs of obligation

The overall distribution of variants of verbs of obligation in the first and second datasets obtained from our informants is shown in Table 4 and Figure 1. HAVE TO is the most used verb by informants from all three countries. However, for

9. Since the figures in the table have been rounded to one decimal place, some of the subtotal scores do not match the added individual values.

10. People who belong to this network (5) overlap with those who belong to one of the four networks above (1, 2, 3 and 4). The network scores in brackets, therefore, are not included in the subtotals.

11. People who belong to this network (8) overlap with those who belong to one of the two networks above (6 and 7). The network scores in brackets, therefore, are not included in the subtotals.

12. We recognise, of course, that patterns of verb choice vary geographically within, for example, the British Isles (see Tagliamonte 2013; Tagliamonte & Smith 2006). We do not present a country-internal investigation of variation here for a number of reasons. Firstly, our speakers
the English and New Zealand informants, HAVE GOT TO is next most popular, while for the Americans it is GOT TO. Notice the low rate of MUST for all informants (cf. Krug 2000), and the low rate of GOT TO for the English and New Zealand informants. The distribution of variants across the three nationalities roughly corresponds to what previous studies have found (e.g. Tagliamonte 2004: 41–42, 2013: 139). If we compare our base-line figures for the different nationalities with those in Collins (2005: 256, 261), we find that the levels of HAVE GOT TO in our data are higher for the Americans and New Zealanders than in Collins’ analysis, though roughly similar for the English. Nevertheless, the fundamental difference between English from England and the United States (US) is replicated in our results – HAVE GOT TO is more common in England (and New Zealand) than the US, HAVE TO higher in the US than England and New Zealand.

Table 4. Overall distribution of variants of verbs of obligation

<table>
<thead>
<tr>
<th>Informants Data</th>
<th>MUST</th>
<th>HAVE TO</th>
<th>HAVE GOT TO</th>
<th>GOT TO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>English 1st dataset</td>
<td>8</td>
<td>9.8</td>
<td>51</td>
<td>62.2</td>
<td>20</td>
</tr>
<tr>
<td>2nd dataset</td>
<td>2</td>
<td>1.7*</td>
<td>56</td>
<td>48.3</td>
<td>52</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>5.1</td>
<td>107</td>
<td>54.0</td>
<td>72</td>
</tr>
<tr>
<td>American 1st dataset</td>
<td>2</td>
<td>3.2</td>
<td>51</td>
<td>81.0</td>
<td>0</td>
</tr>
<tr>
<td>2nd dataset</td>
<td>1</td>
<td>1.7</td>
<td>43</td>
<td>71.1</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>2.4</td>
<td>94</td>
<td>76.4</td>
<td>3</td>
</tr>
<tr>
<td>NZ 1st dataset</td>
<td>4</td>
<td>4.9</td>
<td>43</td>
<td>52.4</td>
<td>27</td>
</tr>
<tr>
<td>2nd dataset</td>
<td>4</td>
<td>5.1</td>
<td>50</td>
<td>64.1</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>5.0</td>
<td>93</td>
<td>58.1</td>
<td>49</td>
</tr>
<tr>
<td>Total 1st dataset</td>
<td>14</td>
<td>6.2</td>
<td>145</td>
<td>63.9</td>
<td>47</td>
</tr>
<tr>
<td>2nd dataset</td>
<td>7</td>
<td>2.8</td>
<td>149</td>
<td>58.7</td>
<td>77</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>4.4</td>
<td>294</td>
<td>61.1</td>
<td>124</td>
</tr>
</tbody>
</table>

Pearson Chi-Square (2-sided):
* significant at $P < .05$,
** significant at $P < .01$.

The use of HAVE GOT TO for the English significantly increases and the use of HAVE TO noticeably decreases from the first dataset to the second dataset. This come from many different parts of each country, with few places represented by more than one person. Secondly, there would be too few tokens per place (= per speaker) to conduct a meaningful analysis. Thirdly, since the types of speakers investigated in earlier geographical studies (e.g. Tagliamonte & Smith 2006) are very different from those in our study, a comparison would anyway bear little validity.
is perhaps not what we might have expected if convergence was taking place. We return to this finding later.

Figure 1. Overall distribution of obligation verbs of 1st and 2nd datasets

5.2 Linguistic constraint analysis: Types of obligation

In the literature (e.g. Tagliamonte 2004: 43–53, 2013: 142–145; Tagliamonte & D’Arcy 2007: 75–81), a number of robust linguistic constraints have been found on variation in verbs of obligation. Given the strength and consistency of obligation source as a constraint across the literature, we examine here how this factor shapes variant choice in our data. For now, the first and second datasets are combined for the constraint analysis.

Table 5 and Figure 2 summarise the distribution of variants according to the types of obligation, namely objective and subjective as shown in examples (5) and (6) above respectively. For all three country groups, HAVE TO is more likely to be used in objective rather than subjective obligation, a finding we have come to expect from the literature (Tagliamonte 2013: 142–145; Tagliamonte & D’Arcy 2007: 76–77; Tagliamonte & Smith 2006: 362).
Table 5. Distributions of variants of verbs of obligation according to the obligation type

<table>
<thead>
<tr>
<th>Informants</th>
<th>Obligation type</th>
<th>MUST</th>
<th>HAVE TO</th>
<th>HAVE GOT TO</th>
<th>GOT TO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>objective</td>
<td>0</td>
<td>0.0</td>
<td>60</td>
<td>67.4</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>subjective</td>
<td>10</td>
<td>9.2</td>
<td>47</td>
<td>43.1</td>
<td>45</td>
</tr>
<tr>
<td>American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>objective</td>
<td>1</td>
<td>1.4</td>
<td>63</td>
<td>86.3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>subjective</td>
<td>2</td>
<td>4.0</td>
<td>31</td>
<td>62.0</td>
<td>1</td>
</tr>
<tr>
<td>NZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>objective</td>
<td>2</td>
<td>2.2</td>
<td>60</td>
<td>65.2</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>subjective</td>
<td>6</td>
<td>8.8</td>
<td>33</td>
<td>48.5</td>
<td>23</td>
</tr>
</tbody>
</table>

Figure 2. Distribution of obligation verbs according to the obligation type

5.3 Impact of the speaker’s social networks on variation

This section examines the correlation between the informants’ social networks and their variant choice. In this section, the percentage scores of individual informants are used for the statistical analyses. Pearson correlation analysis was used first to describe the relationship between the percentage score of each variant and social network index scores of each speaker. Only the variants that showed statistically significant correlations with particular social networks have been chosen to perform further statistical tests. Multiple regression analysis (stepwise method) was carried out to test the strength of the relationship of the variables in question. It is thus possible to predict which of the social networks is mostly influencing the choice of variant. We present here some interesting findings among the English
and American informants. There are no statistically significant correlations found among the New Zealand informants.

Table 6 shows the results of the multiple regression analyses for the informants from England. When we examined the relationship between social network and people’s changes in variant distribution from the first dataset to the second, no significant correlations were found. We then examined the relationship between social network and the patterns in the second dataset alone. First, having a strong “North American social network” has a negative impact on the use of HAVE GOT TO by the English informants. In other words, those with strong social networks with North Americans use less HAVE GOT TO than those who do not, as shown in Figure 3. Second, a strong “network with native speakers of English” has a positive impact on their use of HAVE TO for the English. This means that those with strong social networks with native speakers of English generally use more HAVE TO than those who do not, as shown in Figure 4. Those who do not have strong social networks with native speakers of English tend to use HAVE GOT TO more than those who do, although this tendency is not statistically significant. These findings above may imply that the network strength with North Americans for

<table>
<thead>
<tr>
<th>Variant</th>
<th>Predictor variable</th>
<th>Beta</th>
<th>p</th>
<th>Adjusted R²; F; Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAVE GOT TO in the 2nd dataset</td>
<td>North American network</td>
<td>−.568</td>
<td>.027</td>
<td>Adjusted R² = .271; F₁,₁₃ = 6.199; p = .027</td>
</tr>
<tr>
<td>HAVE TO in the 2nd dataset</td>
<td>Network with native speakers of English</td>
<td>.572</td>
<td>.016</td>
<td>Adjusted R² = .275; F₁,₁₃ = 6.323; p = .026</td>
</tr>
</tbody>
</table>

Figure 3. Correlation between the use of HAVE GOT TO in the second dataset and “North American network” among informants from England
the English informants has an impact on the network strength with native speak-
ers of English as a whole, and, therefore, influences, at least partially, their choice
between HAVE TO and HAVE GOT TO.

![Figure 4. Correlation between the use of HAVE TO in the second dataset and “network with native speakers of English” among informants from England](image)

Table 7 shows the results of the multiple regression analyses for the American informants. First, having a strong “network with English teachers” has a negative impact on people’s changes in the HAVE TO distribution from the first dataset to the second by the American informants. In other words, those with strong social networks with English teachers (native speakers of English and Japanese) tend to decrease the use of HAVE TO after a year and those with weak social networks with English teachers tend to increase the use of HAVE TO after a year, as shown in Figure 5. Second, a strong “network with non-native speakers of English” has a negative impact on their use of HAVE TO after a year. This means that those with strong social networks with non-native speakers of English use less HAVE TO than those who do not, as shown in Figure 6. Those with strong social networks with non-native speakers of English tend to use GOT TO more than those who do not, although this tendency is not supported by statistics.

Table 7. Multiple regression for the American informants

<table>
<thead>
<tr>
<th>Variant</th>
<th>Predictor variable</th>
<th>Beta</th>
<th>p</th>
<th>Adjusted R²; F; Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in the use of HAVE TO</td>
<td>Network with English teachers</td>
<td>- .768</td>
<td>.006</td>
<td>Adjusted $R^2 = .544$; $F_{1,9} = 12.930; p = .006$</td>
</tr>
<tr>
<td>HAVE TO in the 2nd dataset</td>
<td>Network with non-native speakers of English</td>
<td>- .701</td>
<td>.027</td>
<td>Adjusted $R^2 = .434$; $F_{1,9} = 8.672; p = .016$</td>
</tr>
</tbody>
</table>
6. Discussion and conclusion

Firstly we found that in the year between the first dataset and the second dataset, our speakers, both American and English, decreased their use of HAVE TO. For the English, there was an increase in the use of HAVE GOT TO and for the Americans, a small increase in both HAVE GOT TO and GOT TO. It would be very hard indeed to claim that there is convergence in progress here – while the Americans use a higher proportion of the form most associated with the English
HAVE GOT TO, there is a bigger shift towards GOT TO, a form the English barely use at all. Meanwhile the English are quantitatively diverging from the Americans, increasing their use of HAVE GOT TO by almost 20%.

Although they did not account for the shifts across the year, social network strengths were not totally void of importance. Those English who do have strong network ties with Americans tend to use HAVE GOT TO less than those who do not, and those who have strong ties with native speakers in general (as opposed to ties with Japanese, or few ties at all) tend to use more HAVE TO.

How do we account for this failure to converge? One might argue that, for this variable, perhaps for grammatical variables in general, much longer-term contact is needed for convergence to occur, longer than is needed for at least some phonological forms to demonstrate convergence (cf. Thomason & Kaufmann 1988). If this is the case, the methodological conundrum of the intermediate stage of dialect contact shows itself to be especially problematic, given the practical difficulties of collecting data from long-term panel studies, especially in contexts of empirically controllable contact.

Secondly, how do we explain the quite considerable and counterintuitive shift by the English, away from HAVE TO and towards HAVE GOT TO? One might argue that there are methodological reasons: Given claimed associations between the use of HAVE GOT TO and the spoken, informal language (e.g. Smith 2003: 259; Coates 1983: 57), it might be possible to argue that during the first recording session, when both the research process was new and the interlocutor not so well known, the context might trigger fewer informal forms, but a year later, familiar with the researcher, the interlocutor and the research process, the informants might feel more relaxed. This is possible, of course, but given the very informal conversations and the sometimes quite delicate topics raised by the informants in the first recording session, this seems to us unlikely.

Perhaps, though, the structure of the data could help us account for these findings. Earlier, we saw that obligation type/source as a linguistic constraint patterned as one might expect from the existing literature. Let us look at the data obtained from the English informants again. One might find an increase in the use of HAVE GOT TO if the number of tokens in linguistic environments conducive to the use of that variant makes up a larger proportion of the total in the second dataset than in the first. This is, however, not the case with our data for the English informants. As shown in Table 8, the proportions of objective and subjective obligation are almost equal for the first and second datasets. Within the conducive set of

---

13. We are aware of the relatively low number of tokens in each cell when the data are divided into nationality groups, first and second datasets and linguistic category groups as shown in Table 8. This data sparseness, however, does not affect our findings presented in Chapter 5.
subjective obligation tokens, the rate of HAVE GOT TO use has, nevertheless, increased across the two datasets by over 30%, at the expense not only of HAVE TO but also of MUST. There is, therefore, an increase in HAVE GOT TO despite these even token distributions, and, it seems, the sharp increase in HAVE GOT TO is disproportionately found in this linguistic context – the increase of HAVE GOT TO in objective obligation contexts only increased by 9%.

Table 8. Distributions of variants of verbs of obligation in English informants according to the obligation type

<table>
<thead>
<tr>
<th>Data</th>
<th>Obligation type</th>
<th>MUST</th>
<th>HAVE TO</th>
<th>HAVE GOT TO</th>
<th>GOT TO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>1st dataset</td>
<td>objective</td>
<td>0</td>
<td>0.0</td>
<td>27</td>
<td>75.0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>subjective</td>
<td>8</td>
<td>17.4</td>
<td>24</td>
<td>52.2</td>
<td>11</td>
</tr>
<tr>
<td>2nd dataset</td>
<td>objective</td>
<td>0</td>
<td>0.0</td>
<td>33</td>
<td>62.3</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>subjective</td>
<td>2</td>
<td>3.2*</td>
<td>23</td>
<td>36.5</td>
<td>34</td>
</tr>
</tbody>
</table>

Pearson Chi-Square (2-sided):
* significant at $P < .05$,
** significant at $P < .01$.

Relatively speaking, variation in verbs of obligation is not salient when compared with variables with significantly used non-standard variants. We can but speculate at this point about whether our results show that grammatical convergence among adults generally is extremely slow (or non-existent) (and slower, evidence suggests, than convergence for some phonological variables); that grammatical convergence for relatively non-salient features is slow, but can act at a different speed for more salient grammatical variables; or that this variable simply behaves unlike others in this regard. Given, furthermore, the numerical dominance of North Americans among the English teacher population of Japan, the shift towards HAVE GOT TO by all speaker groups represents a rather unusual diachronic adoption of a minority linguistic variant in a dialect contact situation, a variant which is, nevertheless, embedded in the grammar in ways which conform to the expectations of the extensive existing literature. Further research on the behaviour of grammatical variables in contexts of intermediate duration of contact is needed, both to further explore the process by which grammatical patterns of variation evolve when distinct varieties come into prolonged contact, and to assess the extent to which grammatical and phonological convergence mechanically operate in the same way.
Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 25284082.

References


CHAPTER 2

Patterns of linguistic globalization
Integrating typological profiles and questionnaire data

Manfred Krug, Ole Schützler and Valentin Werner
University of Bamberg

In this paper, we present ways of relating the broad scope of typologically moti-
vated approaches to variationist questionnaire data. We explore how descriptive
and inferential statistics (such as linear regression) and exploratory techniques
(such as aggregative analyses) can be combined for a more holistic investiga-
tion of variationist questionnaire data on lexical choices from British, Maltese,
American and Puerto Rican English. Our analyses show that raters from the
British English(-influenced) and the American English(-influenced) sphere form
distinct clusters. Adopting a more fine-grained perspective, we find evidence
both for the actual existence of four distinct regional varieties and for globalizing
tendencies. Based on our results, we further argue that variety-internal variation
is often motivated lexically rather than socially.

Keywords: globalization, cluster analysis, questionnaires, regression models,
varieties of English

1. Introduction

Linguistics has recently seen an increasing interest in aggregative methods (e.g.
Szmrecsanyi & Wolk 2011; Kortmann & Wolk 2012; McMahon & Maguire
2013; Werner 2014; Fuchs & Gut 2016) as well as a continuing interest in World
Englishes and the typological profiling of varieties and structures. Both trends are
reflected in the publication of atlases and handbooks – in online and printed for-
mats – such as The World Atlas of Language Structures Online (WALS, Dryer &
Haspelmath 2013), The Atlas of Pidgin and Creole Language Structures (APiCS,
Michaelis et al. 2013), the Handbook of Varieties of English (Kortmann et al.
2004) and The Electronic World Atlas of Varieties of English (eWAVE, Kortmann
& Lunkenheimer 2013). The present chapter will introduce ways of linking
typologically motivated approaches with a variationist questionnaire approach (cf. Kortmann & Lunkenheimer 2012; Krug & Rosen 2012; Krug & Sell 2013). By zooming in from the typological to the sociolinguistic and language-internal perspective, we will show how a combination of well-established statistical tools (e.g. mixed-effects regression) and exploratory techniques and associated visualizations (e.g. phenograms) can be fruitfully used for a more holistic analysis of variationist questionnaire data spanning several varieties of English.

The present study is based on data collected from 2008 to 2012, using the ‘Bamberg questionnaire for lexical and morphosyntactic variation in English’ (see Krug & Sell 2013 for detail). Material was collected during field trips to several English-speaking countries (see below) by researchers and student assistants from the University of Bamberg. The material was subsequently automatically digitized and, after intensive manual post-editing, stored. This project has produced a fair amount of data for several high- and low-contact varieties of English from locations including Gibraltar, Malta, the Channel Islands, Puerto Rico, Wales, England, Scotland, the USA and Australia. For this chapter, 68 lexical binaries known to differ between British (BrE) and American English (AmE) are inspected (e.g. anticlockwise – counterclockwise; tap – faucet). We highlight two second-language varieties, Malta (219 raters) and Puerto Rico (395 raters), while presenting AmE and BrE data (25 and 14 raters, respectively) as reference points. For each lexical pair, therefore, a total of more than 600 US-American, British, Puerto Rican (PRE) and Maltese English (MaltE) informants have rated their personal preferences for one form over the other on a five-point Likert-scale (cf. Appendix; Krug & Sell 2013: 77–78; Wray & Bloomer 2012: 168). Ratings of the 68 lexical binaries are used to calculate a distance matrix, in which the similarity of any possible two raters is quantified.

In Section 2, we show that raters from the British English(-influenced) and the American English(-influenced) sphere form distinct clusters. Not surprisingly, American and Puerto Rican English raters, as well as British and Maltese English raters, form one cluster each. We find that due to historical and political factors the patterns for lexical choices reflect, inter alia, exonormative influence of American on Puerto Rican English and British on Maltese English. More interestingly, the distance between Puerto Rican and Maltese English is smaller than

1. Of course, the expressions binaries and British vs. American English usage are simplifications, which are used here for expository clarity only. Some items have more than two variants, e.g. dummy – pacifier – soother (or compare X with/to/and Y). In our questionnaire, raters can add comments regarding their own preference and alternative terms in each case. Similarly, we simplistically use BrE (or AmE, as the case may be) when we refer to more British (e.g. backwards vs. backward), exclusively (e.g. -isation spellings) or traditionally British terms (e.g. lorry vs. truck).
that between the two L1 reference varieties, which we interpret as one form of globalization (see Section 3.4 for a fuller discussion). Most interestingly, raters of different varieties overlap only minimally, which reveals strong regional contact and normative factors.

When zooming in on the lexical items rather than individual raters (Section 3), we find exactly the same distinct regional clusters: one BrE(-influenced) and one AmE(-influenced). Importantly, however, we also find statistically significant differences between BrE and MaltE on the one hand, and between AmE and PRE, on the other. This presents unambiguous evidence for the actual existence of four distinct regional varieties rather than for the existence of just two major spheres ([BrE + MaltE] vs. [AmE + PRE]). This is a point worth noting because, especially in Phase III (nativization) of Schneider’s dynamic model, L2 varieties are sometimes claimed not to exist at all, or its features are criticized by conservative speakers adhering to the external norm (cf. the so-called ‘complaint tradition’; Schneider 2007: 56).

In a final step, the Maltese and Puerto Rican data are analyzed using linear regression models. We discuss stratifications of lexical choices along the social dimensions of age and gender, and the relationship between globalization and stable vs. changing usage conventions in different regional varieties. With 20 (in MaltE) and 22 (in PRE) of the 68 binaries responding significantly to the factor age, apparent-time distributions figure prominently in describing our data and in identifying aspects of globalization and language change in progress. In addition, we identify items that are more clearly preferred in their respective American or British variant, while for others a regional preference is less pronounced. We argue that, overall in our data, variety-internal variation is often motivated lexically rather than socially.

2. The macro-perspective: Aggregative analysis

After the collection and digitization of the questionnaire ratings, it is useful to obtain an overview of the general structure of the data. This is a crucial step towards the generation of hypotheses as it informs the further assessment of which

---

2. In this study, we use the uncontroversial term L1 for BrE and AmE. English is an official language in both Malta and Puerto Rico. But English in Malta is used in more domains (including higher education) and is more widely used as a (typically second) native language in Malta than in Puerto Rico (not counting return-migrants). In particular for the Puerto Rican context, therefore, some linguists might prefer the term EFL. By using the term L2 for both Maltese and Puerto Rican English, we embrace a wide definition of L2 as a second language that is used in addition to another (in our case dominant) native language, i.e. Maltese and Spanish, respectively.
categories (in our case: varieties, raters, items, etc.) deserve closer attention. In this regard, exploratory data analysis can help to overcome some of the limitations of the human mind in deducing patterns from large sets of multidimensional quantitative data. In this section, we map relationships between categories (i.e. raters) and thus present a method for tapping into the wealth of information offered by our questionnaire data. The exploratory approach we develop is

1. **data-driven**, as we do not make any prior assumptions that inform the analysis about associations between categories,
2. **aggregative**, in that we conduct a joint analysis of various measurements (cf. Wälchli & Szmrecsanyi 2014: 1); that is, acceptability ratings for different lexical items across different raters, and
3. **visual**, since it exploits the opportunities offered by a graphical representation, and structure(s) in the data can be identified literally at a glance.3

Thus, we can test the similarity and dissimilarity of individual raters across aggregated items. In addition, depending on the outcome of the analysis, we are in a position to make generalizations based on results for individual raters and items, and can determine whether associations and categorizations (e.g. along variety lines) are latent in the data.

### 2.1 Preparing and transforming the data

The starting point for our investigation are the raw questionnaire data, available in spreadsheet format. Each column is headed by an anonymized unique rater ID, followed by the ratings for the individual items covered in the lexical part (ranging from −2 for exclusively AmE usage to +2 for exclusively BrE usage). Note that the spreadsheets also contain the related sociolinguistic information for each rater as well as ratings for grammatical items. Neither of these is genuinely part of the present analysis as such, but may be used at a later point (see Section 2.2). As our first focus lies on the calculation of distances between raters based on their ratings of the 68 lexical items (so individual lexical items do not play a part here, but see Section 3), we can reduce the structure of the data considerably. The simple layout of the resulting file, which is saved in a comma-separated format (CSV) to make it readable for the statistical software that calculates the distance matrices (see below), is displayed in Table 1. It only contains the rater IDs as headers and the ratings for the individual items row by row.

---

3. From the perspective of cognitive psychology, it has been shown that people in general benefit from graphs (in contrast to tables) as regards the understanding of complex relationships between data points (Braithwaite & Goldstone 2013: 1928).
Table 1. Structure of the input file for the calculation of distance matrices

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating item 1</td>
<td>Rating item 1</td>
<td>Rating item 1</td>
<td>Rating item 1</td>
</tr>
<tr>
<td>Rating item 2</td>
<td>Rating item 2</td>
<td>Rating item 2</td>
<td>Rating item 2</td>
</tr>
<tr>
<td>Rating item 3</td>
<td>Rating item 3</td>
<td>Rating item 3</td>
<td>Rating item 3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Rating item 68</td>
<td>Rating item 68</td>
<td>Rating item 68</td>
<td>Rating item 68</td>
</tr>
</tbody>
</table>

Against the backdrop of this structural layout and before we proceed to the actual calculations, we present a brief outline of the general characteristics of the type of aggregative analysis (AA) used as one approach to multidimensional data. It is evident that it is easy to compare two or more categories for one value, such as rater A vs. rater B vs. rater C, etc. for their ratings for item 2 (for instance through the calculation of a mean rating for the relevant row and an assessment of the dispersion), and an identification of differences between categories one at a time is straightforward. However, if we have to deal with an extended number of dimensions along which the categories can be compared (that is, rating values for 68 different items in our case), complexity increases markedly. To briefly illustrate this: Rater A may share 23 values with rater B but may differ in 45 values from her, while he may differ from rater C in 34 other values, etc. Dimensions of comparison multiply as more categories (items) are included.

The initial step is the computation of a distance matrix (alternatively referred to as “(dis)similarity matrix” or “difference matrix”). To this end, all values (rows) for all categories (columns) are compared pairwise, and the correlation between each category is calculated. The correlation value serves as the numerical input for the computation of the overall distance measure between categories, \(^4\) and we obtain a distance matrix from which we can read the overall distance between all possible pairs of categories (i.e. between raters, in our case). Its layout is shown in Table 2.

Table 2. Structure of the distance matrix (adapted from Fuchs & Gut 2016)

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>Overall distance A-B</td>
<td>Overall distance A-C</td>
<td>Overall distance A-...</td>
</tr>
<tr>
<td>Rater B</td>
<td>0</td>
<td>Overall distance B-C</td>
<td>Overall distance B-...</td>
</tr>
<tr>
<td>Rater C</td>
<td>0</td>
<td>0</td>
<td>Overall distance C-...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\(^4\) Other measures may be used for different types of data. As we deal with ordinal variables derived from the questionnaire rating scale, a rank-based correlation seemed adequate.
Note that for the input file of the matrix, values of all types (categorical and continuous) can be used in principle (see, e.g., Kortmann & Wolk 2012; Werner 2014; see Fuchs & Gut 2016 for an accessible step-by-step description of how a distance matrix can be calculated for continuous data). Therefore, AA is able to deal with data of different kinds (relative or absolute frequencies derived from corpus data, categorical expert ratings, questionnaire data, etc.), as long as the categories (varieties, registers, text types, raters, etc.) are comparable. With the distance matrix in hand we are now in a position to establish an intuitive measure of (dis-)similarity between the categories (raters) and thus, to identify latent structure in the data. In addition, the distance matrix facilitates the comparison with multiple neighbours. All this is usually not directly possible relying on manual analysis due to the complexity of multidimensional data.

For the reasons provided above, it is evident that we can exploit the full potential of AA only if effective visualization techniques are available, as we quickly reach the limits of what can be inferred from tabulated numerical data, particularly if many categories are compared. Ideally, the n-dimensional space that was our point of departure (a potentially infinite number of raters multiplied by the 68 ratings) will be reduced to a two-dimensional graphical display, and the perceptual complexity will substantially decrease.

The distance matrices generated may serve as input either for hierarchical cluster analysis, with bifurcating tree-like dendrograms as visual output, or non-hierarchical phenograms (also known as “phylogenetic networks” or “NeighborNets”), both illustrated in Figure 1. As the alternative labels imply, the latter type is an unrooted network, where differences between categories, which are positioned at the nodes of this representation (see below), are shown in terms of distances to all other categories. In both types, possible groupings or associations between categories may emerge.

This type of visualization originated in evolutionary biology (Huson & Bryant 2006) and is used there to chart relationships between bacteria and species, for instance. Due to its versatility, it has been used as an exploratory tool in various areas of linguistics. Among recently published work we find applications in the fields of dialectometry (Szmrecsanyi 2011; Szmrecsanyi & Wolk 2011) and typology (Wälchli 2014), for the study of register differences in corpus data (Werner 2014). A reviewer has pointed out that multidimensional scaling (MDS) is another technique that could be used. For reasons of space, and as MDS can be viewed as a fairly established approach (see, e.g., Kruskal & Wish 1978; Croft & Poole 2008; Szmrecsanyi 2011), we restrict ourselves to an exploration of the potential of non-hierarchical phenograms, only briefly considering hierarchical cluster analysis, which arguably is restricted in its exploratory function compared to non-hierarchical analyses (see below).
This non-hierarchical mode of visualization has two main advantages compared to hierarchical (“tree-like”) representations. First, we do not have to rely on the (for our purposes impractical) assumption of a “common origin” from which all further categorizations derive. And second, we can determine differences also between categories that would appear in a single cluster in a hierarchical cluster analysis (due to the neighbour-joining approach used there). As a result, we obtain a more fine-grained picture (Werner 2014: 132; see also Fuchs & Gut 2016). As these representations may become quite complex, it is important to note that the distance between two categories is always the shortest possible one along the edges of the graph (e.g. as indicated by the bold edges in Figure 2, connecting raters A and E). The actual distance between two categories in the graph corresponds to the distance measure as provided in the distance matrix.

After these more general aspects, some remarks on the tools used for the present study are in order. With R (R Core Team 2014), we computed a distance matrix,
in which the distance of any possible two raters was quantified. Starting from the file containing the raw rating values (see above) we first calculated the Spearman rank correlation between all raters and employed the `dist` method (default settings; Euclidean distance) from the `stats` package.

The choice of Euclidean distance was informed by both practical and theoretical considerations. As to the former, it has been shown to be a convenient and reliable type of measure in other studies and it is established as a “theory-neutral” (Szmrecsanyi 2011: 54) standard distance measure in linguistics. From a theoretical point of view, it may be argued that Euclidean distance is liable to unwanted weighting effects between smaller and bigger frequency differences for individual items (Szmrecsanyi 2011: 54). However, for our data, with its inherent normalization within the range of the five-point Likert-scale (see above), these effects do not play a role. We therefore submit that Euclidean distance is particularly suited for the present investigation. The R output is a distance matrix in a comma-separated format (CSV). After a number of formal adaptations with the help of a purpose-built Python script for the Notepad++ editor, this matrix is loaded into SplitsTree (Huson & Bryant 2006; splitstree.org), where various formatting options for the network representations and an export function for the visualizations are available (we use the default “equal angle” method). What these phenograms look like and why they help us discover patterns in our questionnaire data is shown in the next section.

---

6. There is no internal weighting of the 68 items, so that all ratings contribute equally to the distance measure.

7. Thanks are due to Fabian Vetter for creating the template.
2.2 Applications

As stated, the ratings of the 68 lexical binaries are used to calculate a similarity matrix in which the similarity of any possible two raters is quantified; each node represents one rater. We include (a wealth of) Maltese and Puerto Rican raters as well as (fewer) British and American raters as reference points. For clarity of exposition and due to the methodological focus of this section (aiming to show potential applications), we use a reduced sample of raters \(n = 40\) each for MaltE and PRE, which are randomly chosen from the full set; reference varieties: \(n = 14\) for BrE, \(n = 25\) for AmE). This results in a reduced number of categories that are compared and a reduced number of nodes in the visualizations (see Krug, Schützler & Werner (in prep.) for in-depth data analysis and interpretation of an extended dataset). In the following, we consider a number of graphical representations and their use for an exploration of the data. These allow us to spot anomalies and to identify general trends and structures.

A first application of exploratory AA is the identification of outliers. To this end, we generate a phenogram according to the procedure described in Section 2.1. This rather complex visualization is shown as Figure 3. The general pattern that emerges from Figure 3 is that raters from the British English(-influenced) and the American English(-influenced) sphere form two distinct groups. AmE and PRE raters (left hand side of Figure 3), as well as BrE and MaltE raters (right hand side of Figure 3), form one grouping each. We also find a less crowded area in between, which nicely illustrates the distance between the two broad clusters.

However, we also find raters that do not follow the general trends. One example is rater 59 (MaltE), whose node is highlighted by large bold print in Figure 3. Surprisingly, this data point can be found among the AmE/PRE raters, which suggests that this respondent’s ratings closely match those of the raters from this sphere. At this point, the available sociolinguistic information comes into play to provide a possible motivation for the unexpected pattern. A look at the variable “location(s) where the rater lived” reveals that rater 59, being born in Malta, having lived there for more than 40 years, and currently residing there, has in fact recently spent a three-year period in Canada and the US. This finding provides the most likely explanation for the “Americanized” patterns of this particular rater.

It is evident that the individual researchers – depending on their aims – must decide whether findings such as the one illustrated have an impact on their research agenda. For instance, outliers may be excluded from further analysis as they may (for various reasons) not fit the criteria for the target population as originally specified. At any rate, visualizations as presented above can be viewed as a helpful device to identify anomalies in the data and may, due to their accessible and intuitive nature, serve to inform decisions on how to react appropriately on a
Figure 3. Phenogram (NeighborNet) for the example dataset (lexical questionnaire ratings)
case-by-case basis. This holds true in particular if adequate sociolinguistic information is available to which we can refer.

Some analysts might assume that outliers could also be identified by simply considering the numerical average data rating of the respondents. However, even though the overall average of one respondent across the 68 items may diverge from the average across all respondents from this variety, this does not mean that there is no overlap between this respondent and the other respondents of the same variety at all. What is more, merely looking at overall averages involves introducing another (potentially unwanted) level of abstraction and may therefore lead to an (equally potentially unwanted) exclusion of respondents a priori. The AA approach, we believe, better mirrors the actual ratings for individual items and respondents and therefore allows a more data-driven, fine-grained and thus more accurate filtering.

A second application of the approach is to assess the dispersion of the raters. For the related phenogram presented in Figure 4, we rely on the same dataset as above, with two minor modifications. The outlier, rater 59 (MaltE), is excluded and we add the category “average rater” for each of the four varieties discussed. The data point “average rater” (AVER_*) is a meta-category that contains the average values for each of the 68 items across the collective set of speakers for each variety. Note that this change is the cause for the slightly different shapes of Figures 3 and 4.

This representation allows us to draw a number of inferences and conclusions. First, with the location of the average raters, we are in a position to abstract away from individual raters and can relate the four diatopic varieties to each other. From this perspective, Figure 4 clearly suggests that AVER_US and AVER_GB can be viewed as poles in our study, because they are nearly maximally distant from each other. This is what some might have expected based on earlier claims of “notorious” (Rohdenburg & Schlüter 2009: 1) lexical contrasts between the two varieties. By comparison, the distance between AVER_PR and AVER_MT is smaller, which we interpret as evidence of their greater degree of globalization compared to the L1 varieties. Notice that this is only one of four major types of globalization which we propose and explain in more detail in Section 3.4 below for specific linguistic variables. As our data suggest, however, it can at least partly be transferred to individual raters or language users as well. The scenario encountered here is related to type (3) in Section 3.4, and we suggest that given a large enough set of variables (and one that is not heavily biased towards one of the two major reference varieties), the average L2 user will figure between the major poles of the average BrE and AmE language users. Hence the term globalization here describes the pattern that, in graphical representations of large and varied data sets, L2 varieties will typically be found in the space between BrE and AmE.
Figure 4. Phenogram (NeighborNet) for the adjusted example dataset (lexical questionnaire ratings) with average ratings.
A second (and potentially related) result is that the dispersion of the individual raters around the average data points is much larger for MaltE and PRE when compared to the reference L1 varieties, where many nodes are directly neighbouring. This illustrates the larger amount of internal variation in Puerto Rico and Malta – a point that will be explored in more detail in Section 3 below.8

Notice, however, that for each L1 variety, we can identify two subclusters, of which the bigger one predictably clusters around the respective average data point. Notice also that in the case of the US informants, the smaller subcluster is located considerably more centrally in the phenogram, thus also indicating a more globalized behavior in this subgroup of speakers – another aspect which deserves further attention in future work.

2.3 Limitations

Although all tools used (R, Notepad++, SplitsTree) are open-access and thus available free of cost, and AA is a powerful and highly versatile method for dealing with data of the type described above, some caveats need to be addressed. It goes without saying that the general approach presented is labour- and cost-intensive in terms of (1) travel costs and human resources required for data collection during fieldwork, and (2) human resources required for the manual post-checking of the subsequent digitization.9 However, these are the basic requirements to conduct an AA of the type projected in the first place. First, online questionnaires, which are easier to distribute and manage, are not as reliable and do not reach all social strata. Second, the benefits of additional perspectives (to complement corpus data, for instance) justify and, in our case and our view, at least, outweigh the costs.

Some methodological issues are also worth mentioning. First, and generally, in the application of phenograms, if a great number of nodes are included, representations may become overly complex. This may lead to uncertainty as to whether the complexity of a phenogram impairs its interpretability (and thus conceals

8. Note that data collection for the L1 varieties mainly took place in university contexts, for the L2 varieties in both university contexts and by approaching informants in public and non-public places (e.g. streets, parks, shops, homes for the elderly). Our L1 data should therefore be viewed as somewhat more controlled for social variation, which may also contribute to a greater variety-internal homogeneity. On the other hand, diatopic and stylistic differentiation is greater in L1 varieties than in L2 varieties of English (Schneider 2007) so that this aspect must remain a topic for future investigations. We plan to collect and examine a wider range of social strata also for the L1 varieties, of course.

9. As of December 2015, funding and expenses for the project (which includes the compilation of the Maltese and Puerto Rican components of the International Corpus of English) amounts to approximately €250,000.
existing patterns or structure) or whether the displayed complexity in actual fact reveals the non-existence of patterns in the data. To avoid this dilemma, one has to find a sensible upper limit of categories included and, if necessary, take a sample from the total number. Second, and more specifically, it will not be possible to explain each individual outlier, as the relevant sociolinguistic information may not have been recorded during fieldwork.

In addition, other factors need to be considered that may lead to the exclusion of individual informants or particular data points (for instance, when raters do not concentrate while filling in a questionnaire, which is obvious when an informant from some point onwards always chooses the same column in a questionnaire, or is observed to choose categories fast and randomly).

Lastly, missing values may also occur in the ratings. As the calculation of correlations and similarity matrices relies on complete datasets (see Section 2.1), a policy for compensating for these empty cells has to be developed. For our data, for example, it proved feasible to insert the average value across the raters for one variety for this particular item, if an empty (or non-numerical) cell occurred. Thus, skewing effects could be avoided as far as possible.\textsuperscript{10} Other types of data may naturally require different strategies.

In spite of such caveats, we hope to have shown that AA and its related visualization through phenograms can contribute substantially to the exploration of large multidimensional datasets. It reduces inherent complexity and provides additional insights into the structure of the data that would otherwise remain undiscovered. It is obvious, however, that we need further approaches to obtain a perspective that is as comprehensive as possible. We outline possible avenues in the following section.

\textsuperscript{10} An alternative strategy that throws into relief the difference between the rater and his/her variety is used in Krug, Schützler & Werner in preparation: It multiplies the average of the regional variety for the missing value with the individual rater’s overall score relative to the overall score of the regional variety. In that case, transformation of our data set from \(-2\) to \(+2\) into \(1\) to \(5\) is advisable to avoid multiplication with zero. In addition, the upper and lower limits must be observed even after multiplication. The overall distribution of our raters, however, remains virtually identical, irrespective of the method applied. The findings reported here can therefore be regarded as robust, and methodological skewing seems minimal. It is evident that researchers may want to exclude respondents if the number of empty cells exceeds a certain cut-off point.
3. **Zooming in: Testing individual items and social factors**

If the previous section was concerned with the general clustering of speakers of the different varieties under investigation, the following discussion shifts the focus to the evaluation of individual lexical binaries. Important questions are the following:

1. How strong is the influence of the respective typical British or American form in each of the four varieties?
2. Can speaker preferences be modelled, i.e. are there general sociolinguistic patterns showing that speakers of different age or gender make significantly different choices in lexical usage across our entire sample from the lexicon?
3. Can similar sociolinguistic patterns be detected for individual items?
4. Which items diverge most strongly from the expected local or national standard and can be argued to be heralding globalizing tendencies in the English lexicon?

The following sections illustrate what kinds of insights we can hope to gain by shifting our focus from the global to the particular, or, from the speaker-based to the item-based perspective.

3.1 **Regional means**

In order to generate mean values for all 68 lexical binaries, a mixed-effects linear regression model containing no predictor variables – known as a null-model (Luke 2004: 19) – was applied to the data. The function *lmer* from the R-package *lme4* (Bates et al. 2015) was used, and the model was specified as shown further below in Table 3. The model calculates coefficients for all 68 lexical items, which are equivalent to their mean values. Figure 5 displays the distribution of these values for all four varieties under investigation. The horizontal bar in each plot represents the value of the median, and the symbol $\mu$ additionally indicates the position of the arithmetic mean, whose position reflects the general skew of the data away from the normative pole in the respective variety. The box contains the central 50% of cases and the whiskers embrace the remaining two quartiles excluding outliers. The notches in each plot can be treated as confidence intervals, indicating a significant difference between groups if they do not overlap.

---

11. In fact, using mean ratings for individual items as a diagnostic of general differences would have been an alternative. We nevertheless decided to use the coefficients from mixed-effects models to keep Sections 3.1, 3.2 and 3.4 methodologically homogeneous.

12. Notches indicate the range of $\pm 1.58$ interquartile ranges (box lengths) around each mean, divided by the square root of the number of observations.
Outlier items (in the order of increasing distance from the median) are compare x to y and to licence in British English, whilst, ill, forwards, sport and lorry in Maltese English, book about chemistry, storm in a teacup and subway in Puerto Rican English, forwards, autumn, Anyway..., roundabout, compare x to y, book about chemistry, backwards, subway and typical of in American English. These and other items diverging markedly from expectations are discussed in more detail in Section 3.4 (cf. Figure 8).

![Boxplots of distribution of mean ratings for 68 lexical items in four varieties](image)

**Figure 5.** Boxplots of distribution of mean ratings for 68 lexical items in four varieties

The medians of all four varieties are clearly different. The median rating for Britishness decreases in the following order:

British English > Maltese English > Puerto Rican English > American English

and the large gap between the British-oriented and the American-oriented spheres is in accordance with the major clusters identified in Figure 3 above. The notches in the plot do not overlap between any two varieties, and differences at this level can therefore be regarded as statistically robust. In other words, based on these 68 lexical binaries of the present study, it is statistically justified to speak of four distinct regional varieties, two of which belong to the BrE-oriented (or BrE-influenced) sphere, and the other two of which belong to AmE-oriented (or AmE-influenced) sphere. Even though we did find regional clusters and subclusters in the phenograms focusing on individual speakers (cf. Section 2 above), this result was less unambiguous, and thus this second part of our study represents a substantial step forward in our understanding of the data.

It can be observed that the interquartile range (the length of the box) is greater in L2 varieties, which points to a less uniform behavior of items in these varieties compared to the two standard reference dialects. This is equally true for the
distribution at large, beyond the inter-quartile range. One possible hypothesis would be that the less compact dispersion of ratings in Puerto Rican English will be accompanied (and perhaps caused) by more pronounced inter-speaker differences that are sociolinguistically motivated. Some evidence that this is indeed the case will be provided below (cf. Section 3.3).

3.2 Application of global mixed-effects linear regression models

Again using the function \texttt{lmer} from the R-package \texttt{lme4}, we applied a mixed-effects linear regression model predicting the expected rating of lexical items, including AGE and GENDER as social predictors. For this chapter, we only tested Maltese and Puerto Rican English.

Random effects were specified for rater and lexical items. This is necessary whenever individual observations (in this case ratings) are nested within higher-order groups (in this case raters and lexical classes). Data structures of this type require mixed-effects (or random-effects) models, since they violate the assumption of independent observations, which is fundamental to ordinary least-squares regression, for example (cf. Gries 2013: 333–336).

Table 3 shows the four models that were fitted to the data. The most comprehensive one (Model 3) contains both social predictors and their interaction;\(^\text{13}\) in a less complex model, the interaction term is removed (Model 2); and the most basic models contain only one of the two social predictors (Models 1a and 1b).

<table>
<thead>
<tr>
<th>Model</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1 + AGE + GENDER + AGE*GENDER + (1</td>
</tr>
<tr>
<td>2</td>
<td>1 + AGE + GENDER + (1</td>
</tr>
<tr>
<td>1a</td>
<td>1 + AGE + (1</td>
</tr>
<tr>
<td>1b</td>
<td>1 + GENDER + (1</td>
</tr>
</tbody>
</table>

In both varieties (Maltese and Puerto Rican English), there were no significant effects whatsoever in any of these models. This indicates that at a global level, i.e. across all lexical items, there are no effects of social factors on the ratings that were made. We show in the next section that this does not entail that there are no sociolinguistic patterns in the data, but is due to the fact that a single factor like GENDER may affect different items in different (or even opposite) ways, and that these effects may cancel each other out. In other words, individual lexical binaries

\(^{13}\) The interaction term explores the possibility that AGE has an effect only on men (or women), or, conversely, that GENDER has an effect only on older (or younger) raters.
truly behave as individuals, and their socially conditioned variation will not be captured if a predictor is applied to all items at once.

3.3 Tracing age effects in individual items

For this part of the analysis, an ordinary least-squares linear regression model was applied to the data. This model differs in that it does not make predictions across the entire dataset, but only predicts ratings made for a single lexical pair at a time. Moreover, no random effects for raters or lexical items need to be included, because in each run of the analysis only one lexical item is of interest and each rater only contributes one rating. Other than that, the analysis proceeds along similar lines as those outlined in Section 3.2 for the mixed-effects model. Using the function \texttt{lm} from the R-package \texttt{stats}, which is included in the standard packages that come with R, the first step was to fit a full model including the predictor variables \texttt{AGE}, \texttt{GENDER} and their interaction \texttt{AGE*GENDER} (Model 3). If non-significant, the interaction was removed, yielding a model only containing the two main effects \texttt{AGE} and \texttt{GENDER} (Model 2), one of which was then removed, if necessary, resulting in a model with a single predictor (Models 1a and 1b). The three steps are summarized in Table 4.

Table 4. Ordinary least-squares linear regression models fitted to the data

<table>
<thead>
<tr>
<th>Model</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>\texttt{AGE + GENDER + AGE*GENDER}</td>
</tr>
<tr>
<td>2</td>
<td>\texttt{AGE + GENDER}</td>
</tr>
<tr>
<td>1a</td>
<td>\texttt{AGE}</td>
</tr>
<tr>
<td>1b</td>
<td>\texttt{GENDER}</td>
</tr>
</tbody>
</table>

It was obvious from the single-item linear regression models that raters’ age and gender affects more items in Puerto Rican English than in Maltese English. Table 5 shows that in contrast to Puerto Rican English, none of the more complex models (Models 2 and 3) were significant in Maltese English. Of the simpler models containing only \texttt{AGE} or \texttt{GENDER} as predictors (Models 1a and 1b), an equal number of our binaries (25 each) was significant in both varieties. Consequently, the number of items that did not respond to \texttt{AGE} or \texttt{GENDER} at all is larger in Maltese English; in other words, these sociolinguistic variables have a greater impact overall for our lexical binaries in Puerto Rican English.

The more fine-grained patterns that can emerge from such a set of individual analyses will be illustrated by focusing on those items that responded significantly to \texttt{AGE} only ($n = 20$ for Maltese English and $n = 22$ for Puerto Rican English). This
dimension figures as the single most influential factor in our data set, and – as age-grading is not an important phenomenon in our data set – it is one which has important implications on aspects relating to globalization (cf. Cukor-Avila & Bailey 2013 on apparent-time distributions and their relation to real-time language change). Figure 6 displays an overview of age effects, which are an extrapolation from the output of the linear regression model to two idealized groups of raters, one for which Age was fixed at value 25 (‘younger’), and another for which Age was fixed at value 75 (‘older’).14 The parts above the solid horizontal line in each panel of Figure 6 show those items that seem to be undergoing an apparent-time change towards a more American usage, whereas the bottom parts show those that are developing towards a more British usage. More extreme effects are found towards the top or bottom of each panel.

In a format of this kind, it is possible to identify trends that affect individual lexical pairs or groups. Our first observation is that there is no clear general tendency of change in one or the other direction, that is, moving away from or consolidating the expected influence of American and British English, respectively. In Maltese English, exactly 10 items change in each direction in apparent time, while in Puerto Rican English 13 out of 22 items displaying an apparent-time trend are shifting towards the British pole of the continuum. Items that respond to Age in both varieties are colour and roundabout, both of which are developing towards a more British usage, which might point to a globalizing tendency in the guise of Britishisation. In both cases, however, young MaltE and PRE speakers still opt overwhelmingly for the regionally expected variants; but while roundabout for younger MaltE speakers has become a near-norm, younger PRE speakers are moving towards free variation between roundabout and traffic circle.

In both PRE and MaltE, compare x to y and to licence are developing towards a more American usage, and in both varieties, with scores for 25-year-olds hovering

---

14. We would like to thank Lukas Sönning for assisting with the details of Figures 6 and 8. +2 on the X-axis (“Britishness”) is obtained when every rater reports to exclusively use the (more) British variant; −2 is obtained when every rater reports to exclusively use the (more) American variant.
Figure 6. Age effects in Malta and Puerto Rico: Responses typically expected from 25-year-old raters (‘younger’) and 75-year-old raters (‘older’)

<table>
<thead>
<tr>
<th></th>
<th>younger o</th>
<th>older +</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Puerto Rico</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anyways, ...</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>maths</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>glocalisation</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>to licence</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>typical of</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>localisation</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>compare X and Y</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>autumn</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>football</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>colour</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>launderette</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>rubbish</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>aubergine</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>potato chips</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>roundabout</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>pavement</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>subway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aluminium</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>chips</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>boot</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>windscreen</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>a tap</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Malta</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to licence</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>parcel</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>launderette</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crisps</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>compare X and Y</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>book about chemistry</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>to let</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>ill</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>dummy</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>lorry</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>postman</td>
<td></td>
<td></td>
</tr>
<tr>
<td>roundabout</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cinema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anticlockwise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>colour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>touch wood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>holiday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>centre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jacket potato</td>
<td></td>
<td></td>
</tr>
<tr>
<td>backwards</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Britishness</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
</table>
around zero, there is a trend towards free variation. These items, we believe, display globalization in the guise of Americanization. *Launderette*, by contrast, follows a diametrically opposed trend in PRE and MaltE, in both cases moving away from the expected standard. What unites the three items *compare*, to *licence* and *launderette*, is their trend among younger speakers towards free variation; what separates them is that this trend is not equally pronounced.

Patterns of the kind that we identified in this section, as well as gender-related patterns that we plan to explore more exhaustively in future research, can shed light on what we would like to call globalizing tendencies in lexical usage. That is, a number of items are shifting to a middle-ground position between the two postulated standards. This lack (or loss) of exonormative orientation may well be a characteristic of English L2 varieties more globally and may eventually feed back into L1 usage as well. This issue will be pursued from a slightly different perspective in the following section.

### 3.4 Identifying globalizing lexical items in L1 and L2 Englishes

In this section, we discuss several forms of and possible statistical hints at globalization. Taking as a starting point Trudgill’s (1998) seminal paper on convergence vs. divergence in varieties of English, we propose the following main scenarios for developments in the lexicon (and possibly elsewhere):

1. **Convergence on former British norms**
2. **Convergence on former American norms**
3. **Convergence on free variation between former British and American norms**
4. **Convergence on a norm that is neither a former (standard) British nor former (standard) American norm, but potentially originates in a low-prestige dialect or L2 variety**

Obviously, we understand *convergence* not as a product (‘identity’ across all varieties) but as a process of ‘increased or increasing similarity’. In what follows, we spell out the above four scenarios in some more detail. We then investigate those items that are currently losing their orientation towards either the British or American pole of usage, and discuss the relation of these developments to globalization in L2 varieties of English.

The most uncontroversial cases of globalization are probably scenarios (1) and (2), which figure as the parallel spread of an originally AmE form in varieties that are clearly oriented towards AmE, but also in varieties that are less obviously oriented towards AmE; or vice versa, in the world-wide spread of an erstwhile BrE form. Items that lose their orientation towards either the British or the American pole of usage (and instead converge on free variation between such former poles)
can equally be viewed as tokens of incipient globalization: in this case (scenario 3), globalization results in a (temporary or permanent) loss of orientation, or, put differently, in the loss of a single regional norm.

To trace such patterns in the data, our focus in this section is on the parallel (or non-parallel) behavior in different varieties of English of those items for which the expected national or regional preference is relatively weak: Do lexical items that rank low for Britishness or Americanness in the respective L1 variety follow the same pattern in the corresponding L2 variety? Is there, in other words, evidence for once dominant British variants to be neutralized in British-oriented spheres? Or, by analogy, is there evidence for the inverse scenario, in which a once dominant American variant ceases to dominate in American-oriented spheres? And, finally, as a test case for scenario (3) above, are certain items found in the relatively neutral middle-ground for all four varieties? Scenario (4) cannot be tested quantitatively with the help of the present questionnaire because it does not focus on newly emerging variants.

The first general finding is that the average ratings of items are positively correlated both between BrE and MaltE and between AmE and PRE, as shown in Figure 7, using Pearson’s r. Again we used the coefficients for lexical items from the mixed-effects regression models (see Section 3.1 and Footnote 11).

In the British-oriented sphere (BrE and MaltE), most items cluster in the region where ratings are well above zero or even higher than value 1 for both varieties. In the American-oriented sphere (AmE and PRE), by contrast, most items cluster in the region where ratings are well below 0 or even below −1. In both plots, however, several items do not conform to this pattern, and those will be inspected in more detail in the following paragraphs.

In order to get a more detailed picture and to identify the interesting items in question, Figure 8 plots those 25% of the 68 items in each variety that had the most divergent coefficients (i.e. relatively low values in MaltE and BrE and relatively high values in AmE and PRE). In British English, 18 items are listed, because the items typical of and rubber had exactly the same coefficient and we decided to report both. Figure 8 also explicitly shows the outliers indicated in Figure 5 above. An alternative approach would be to define a threshold either based on absolute average values or on a measure of dispersion (e.g. z-scores).

In total, 35 different items are included among the most divergent 25% in at least one of the four varieties we inspected. While Figure 8 is useful mostly in order to identify which items are affected in each variety, a more regular pattern only emerges when items are tabulated, as shown in Table 6. Here, each lexical item is marked as ‘x’ if it is among the most divergent (‘globalizing’) 25% in the respective variety.
Figure 7. Correlating average ratings of lexical items in L1 and L2 varieties

Coef. (Malta)

Coef. (UK)

r = .644
p = .000

Coef. (Puerto Rico)

Coef. (US)

r = .752
p = .000
This kind of representation shows the regularities that are present in the data more clearly. There are three items that have globalized in all four varieties, 13 items that have globalized in both varieties from the same (BrE-oriented or AmE-oriented) sphere, six items that have globalized in both varieties from the same sphere but which also show this tendency in one variety from the other sphere, and 13 items that have globalized in only one variety. It seems that the normative influence may start to weaken in any variety, L1 or L2, as shown by items listed in the bottom part of Table 6. Nevertheless, the association between the respective L1 and L2/EFL variety in each sphere seems to have quite a strong influence: We believe that the respective L1 variety is involved at some stage in the diffusion process, which is indicated by the fact that there are no items in Table 6 that are only affected in either Maltese English or Puerto Rican English and the L1 variety from the other sphere; that is, the combinations MaltE/AmE and PRE/BrE do not occur. This is a hypothesis that needs to be tested in future research with a focus on the sequence in which patterns emerge.

For some items, the present chapter already provides a relatively detailed picture. For example, in Maltese English to licence, parcel, crisps and compare x to y are included among the globalized items (cf. Table 6) and at the same time show a globalizing apparent-time effect (cf. Figure 6). The same is true of the items subway, a tap and aluminium (cf. Spanish aluminio) in Puerto Rican English, so that for this group of items we can already partly explain the process that led to their globalization. However, all of the approaches presented in this section need to be refined in future research. In addition to average values (or coefficients) and age effects, the variability of items must be taken into account, since it can serve as an indicator of destabilization and thus incipient globalization. Together with a discussion of additional social factors, first and foremost gender, we hope to progress one step further towards a better explanation of the processes that are at work in the lexicon of the varieties included in our study.
**Patterns of linguistic globalization**

Table showing the most globalizing 25% of items in four varieties:

- **United States**
  - typical of
  - backwards
  - book about chemistry
  - compare X and Y
  - roundabout
  - autumn
  - forwards
  - a drop in the ocean
  - ill
  - pavement
  - storm in a teacup
  - sport
  - holiday
  - railway
  - a tap

- **Puerto Rico**
  - subway
  - storm in a teacup
  - backwards
  - typical of
  - a drop in the ocean
  - compare X and Y
  - bicentenary
  - aluminium
  - postman
  - a tap
  - to licence
  - holiday
  - dummy
  - sport
  - cinema

- **Great Britain**
  - rubber
  - realisation
  - forwards
  - ill
  - parcel
  - potato chips
  - to let
  - potato crisps
  - bicentenary
  - jacket potato
  - bookings
  - subway
  - sport
  - book about chemistry
  - whilst
  - compare X and Y
  - to licence

- **Malta**
  - book about chemistry
  - crisps
  - potato crisps
  - aubergine
  - bicentenary
  - to let
  - compare X and Y
  - to licence
  - bookings
  - potato chips
  - jacket potato
  - parcel
  - whilst
  - ill
  - forwards
  - sport
  - lorry

---

**Figure 8.** Most globalizing 25% of items in four varieties
Table 6. Items among the most globalizing 25% in four varieties

<table>
<thead>
<tr>
<th>Variety</th>
<th>GB</th>
<th>MT</th>
<th>PR</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>book about chemistry</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>compare x and y</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>sport</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>bicentenary</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>forwards</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>ill</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>to licence</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>subway</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>typical of</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>bookings</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>parcel</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jacket potato</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>potato chips</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>potato crisps</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to let</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>whilst</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a drop in the ocean</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>a tap</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anyway, …</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>backwards</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>holiday</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>storm in a teacup</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>realisation</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rubber</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>crisps</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>aubergine</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>lorry</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>postman</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cinema</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aluminium</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dummy</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>railway</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>roundabout</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pavement</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>autumn</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Conclusion

We hope to have shown how a questionnaire study using a combination of analytical methodologies can offer a more complete picture of lexical (and, in principle, also grammatical) variation than typological profiles (e.g. based on expert ratings) or questionnaire studies could offer on their own. In the aggregative analysis focusing on individual raters, we found one British-oriented cluster (Britain and Malta) clearly contrasting with one US-oriented cluster (mainland US and Puerto Rico). Regional subclusters were identified and, in particular, a calculation of average raters showed that British and American English indeed form poles in our dataset, between which the L2 varieties of Maltese English and Puerto Rican English are situated (cf. Figure 3 above). Thus, while each L2 variety of English clusters with its reference L1 variety (BrE and AmE, respectively), the two L2s are closer to one another than each of them is to the reference variety of the other L2. We interpret this fact as reflecting stronger effects of linguistic globalization in L2 varieties than in L1 varieties.

The similarity of a reference variety and its contact L2 variety was also in evidence in the analysis focusing on the individual lexical items rather than on the raters (see Figure 5 above). Notably, however, the present data show that there are substantial and statistically robust differences between British, American, Maltese and Puerto Rican English that can support claims as to the distinctness of these varieties – and thus their actual existence in the first place – although we acknowledge the self-evident idealizations involved.

Another robust finding was the wider dispersion of the data and thus the heterogeneity of L2 varieties and their individual raters, irrespective of the methodologies adopted (see Sections 2 and 3 above). This points to a greater homogeneity of L1 varieties, other things being equal. As long as we investigate mainstream speakers of English (and not, say, NORMS), it seems intuitively plausible that the – often exclusive – use of English in daily interaction leads to a more homogeneous language use for lexical items in L1 varieties than the less frequent to sporadic use of English in L2 and EFL varieties.

This study reveals another noteworthy facet of globalization: In our data set, we found that some lexical choices (such as compare, to licence and launderette) move from clearer regional patterns (i.e. reference-variety oriented) among the older cohorts towards free variation among younger cohorts. This, together with, for instance, the variation between roundabout and traffic circle in younger PRE speakers, highlights an interesting phenomenon: While previous studies have usually argued for a dichotomy between Americanization and Britishisation (and sometimes for globalization in either direction, though typically towards American usage; cf. Trudgill 1998; Mair 2006), we have presented evidence that
globalization may figure as a trend towards free variation. Whether or not we are dealing with ephemeral trends that eventually result in the dominance and spread of one variant at the expense of its currently free variant(s) is difficult to predict, though we would assume that free variation is more likely to prevail with subtle spelling differences like *licence* vs. *license* and grammatical variation (as with *compare X to/with or even and Y*) than with concepts such as *roundabout*.

While we have tested for age and gender in this study, we aim to investigate more closely other linguistic and extralinguistic factors in the future, such as education or parents’ native language. The effect of word frequency (Bybee 2010) as a conditioning factor is another promising field of investigation. On a related note, usage domains seem worth investigating as, for instance, household vocabulary appears to be rather stable in Malta (cf. Krug 2015).

We are currently working towards a more flexible database which facilitates the generation of input for statistical calculations. This will help us overcome the limitations of Excel spreadsheets and enable us to inspect the original data sheets with less effort (ideally with a few clicks only), for instance in the identification of outliers. And although the overall patterns are theoretically plausible and appear stable under different methodologies despite comparatively low numbers for BrE and AmE, we plan to include more speakers from reference varieties and other L1 and L2 varieties of English in order to retest and refine our hypotheses and conclusions. Therefore, we hope to substantially enlarge this database in the future, also including learner Englishes (such as German English, Swedish English, Polish English, etc.). Furthermore, we plan to make it accessible to the research community when the digitization of the material is completed.

References


Appendix

Example of Lexicon Questionnaire (Maltese Version), items 1–34

<table>
<thead>
<tr>
<th>Maltese English Questionnaire</th>
<th>Lexical Items</th>
<th>Explanation / Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>a drop in the ocean</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>aluminum</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>anticlockwise</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>eggplant</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>backward</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>bicentenary</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>cookie</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>bookings</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>trunk</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>car park</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>center</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>chemist’s</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>French fries</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>fries</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>cinema</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>color</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>cupboard</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>driver’s license</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>dummy</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>trash can</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>fish fingers</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>soccer</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>forwards</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>globalization</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>glocalisation</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>vacation</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>liberalization</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>baked potato</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>launderette</td>
<td>always use this expression</td>
<td></td>
</tr>
<tr>
<td>potato crisps</td>
<td>always use this expression</td>
<td></td>
</tr>
</tbody>
</table>

I always use this expression
I use this expression more often
I have no preference
I use this expression more often
Never use either expression

(source of the year)

(fruit/vegetable)

(something sweet to eat)

(of a car)

(spelling)

(for clothes)

(kicking game, only goalkeeper uses hands)
Example of Lexicon Questionnaire (Maltese Version), items 35–68

**Maltese English Questionnaire**

**Lexical Items**

<table>
<thead>
<tr>
<th>Lexical Items</th>
<th>Explanation / Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>to licence</td>
<td>to license</td>
</tr>
<tr>
<td>elevator</td>
<td></td>
</tr>
<tr>
<td>localisation</td>
<td></td>
</tr>
<tr>
<td>truck</td>
<td>lorry</td>
</tr>
<tr>
<td>maths</td>
<td>math</td>
</tr>
<tr>
<td>cell phone</td>
<td>mobile phone</td>
</tr>
<tr>
<td>modernisation</td>
<td>modernization</td>
</tr>
<tr>
<td>diapers</td>
<td>nappies</td>
</tr>
<tr>
<td>organisation</td>
<td>organization</td>
</tr>
<tr>
<td>package</td>
<td>parcel</td>
</tr>
<tr>
<td>pavement</td>
<td>sidewalk</td>
</tr>
<tr>
<td>gasoline</td>
<td>petrol</td>
</tr>
<tr>
<td>petrol station</td>
<td>gas station</td>
</tr>
<tr>
<td>mailman</td>
<td>postman</td>
</tr>
<tr>
<td>pushchair</td>
<td>stroller</td>
</tr>
<tr>
<td>railroad</td>
<td>railway</td>
</tr>
<tr>
<td>realisation</td>
<td>realization</td>
</tr>
<tr>
<td>traffic circle</td>
<td>roundabout</td>
</tr>
<tr>
<td>rubber</td>
<td>eraser</td>
</tr>
<tr>
<td>trash</td>
<td>rubbish</td>
</tr>
<tr>
<td>shopping trolley</td>
<td>shopping cart</td>
</tr>
<tr>
<td>sports</td>
<td>sport</td>
</tr>
<tr>
<td>storm in a teacup</td>
<td>tempest in a teapot</td>
</tr>
<tr>
<td>underpass</td>
<td>subway</td>
</tr>
<tr>
<td>to let</td>
<td>for rent</td>
</tr>
<tr>
<td>flashlight</td>
<td>torch</td>
</tr>
<tr>
<td>touch wood</td>
<td>knock on wood</td>
</tr>
<tr>
<td>sneakers</td>
<td>trainers</td>
</tr>
<tr>
<td>whilst</td>
<td>while</td>
</tr>
<tr>
<td>windshield</td>
<td>windscreen</td>
</tr>
<tr>
<td>a book about chemistry</td>
<td>a book on chemistry</td>
</tr>
<tr>
<td>compare X to Y</td>
<td>compare X with Y</td>
</tr>
<tr>
<td>typical of</td>
<td>typical for</td>
</tr>
<tr>
<td>Anyways, ...</td>
<td>Anyway, ...</td>
</tr>
</tbody>
</table>

Informant ID #

Date
Chapter 3

The substitutability and diffusion of want to and wanna in world Englishes

Eugene Green
Boston University

Occurrences of want to and wanna in the extensive database GloWbE have widespread, though disparate, frequencies in varieties of English. This diffusion of want to as consistently greater in frequency than wanna throughout twenty varieties, however, awaits further sampling and analysis. As for substitutability, want to and wanna recur in nearly all environments. Two deterrents to such substitutability, one semantic, the other structural, are due to institutional and spoken practices. In institutional settings, want to in the sense 'obligation' prevails exclusively. In speech want to again prevails over wanna (but not exclusively) in clause final position. One emergent practice finds wanna, uninflected and unrelated to the infinitive marker to, colligated with nouns and noun phrases. This practice is indicative of unforeseen patterns, related to wanna and want to, likely to arise in colloquial English.

Keywords: phonological features, verb phrases, clausal patterns, functional shift, semantic dimensions, pragmatics

1. Introduction

Discussions of standard want to and informal wanna offer analyses of their phonological features, their semantic range, and their possible substitutability in selective grammatical as well as pragmatic contexts. On their phonological features, Krug (2000: 210–211) likens want to and wanna to the pairs going to – gonna, and got to – gotta. The informal variant has two syllables, the first accented, the second vowel mostly schwa. Hudson (2006: 607) reports, from a small sampling, that informants vary their pronunciation of wanna’s unaccented vowel: some have, instead, the

---

1. Many thanks to two anonymous reviewers for their valuable guidance; the shortcomings that remain are my own.
rounded /ə/, others /i/ before a following, initial vowel, others still an intrusive /r/ between vowels. For wanna, omitting /t/ results in consonant cluster reduction. As for the semantic range of want to and wanna, both indicate volition; Krug (2000: 184) also suggests intentions, hypotheticals, obligations, commands, and perhaps probabilities. Examples of intention from the British National Corpus include:

(1) Mind you, they’ll probably wanna do overtime next week!

(2) Yes, I want to go to the United States to experience a different culture. I’m doing Communication Studies which involves advertising and the media. I reckon the US is geared more towards the media. (quoted in Krug 2000: 149)

In (1) mind you and the exclamation mark together pragmatically underscore the speaker’s assertion on the likely intentions attributed to agents. In (2) the opener Yes pragmatically underscores the speaker’s intention. The second and third sentences discursively support this speaker’s intention. Krug’s implicit reliance on pragmatic markers for indicating that intention is a semantic property of wanna and want to prompts further consideration of how examples like his work.

A complementary consideration that Hudson (2006) plumbs is the possibility of interchanging want to and wanna in grammatical contexts. One context that he discusses is the wh- question, Who do you want to/wanna meet? (2006: 604). Although he notes that Who do you wanna meet? is grammatically possible, it occurs nowhere (unlike Who do you want to) in the Corpus of Web-Based Global English (GloWbE).2 Another context that permits possible substitution of want to/wanna involves imperatives. Hudson offers Come on – really want to do it, and you’ll do it!. GloWbE has but few examples of both want to and wanna in imperatives:

(3) You don’t want to be the guy who was left out of the premiere of the best Bond film ever. Neither do you want to watch it in a 50 bob camera copy disc. (KE)

(4) ohh you wanna watch it with the blutac! It might melt through your stick or something! (IE)

Grammatical substitutability includes, not merely categorical differences as in How much paint do you want to/*wanna finish the house (Hudson 2006: 605) but also degrees of diffusion throughout world Englishes. Thus, the GloWbE corpus

---

2. Abbreviations for the twenty variants of English collected in GloWbE appear in this order: US (United States), CA (Canada), GB (Great Britain), IE (Irish English), AU (Australian), NZ (New Zealand), IN (Indian), LK (Sri Lanka), PK (Pakistan), SG (Singapore), MY (Malaysia), PH (Philippines), HK (Hong Kong), ZA (South Africa, NG (Nigeria), GH (Ghana), KE (Kenya), TZ (Tanzania), JM (Jamaica).
contains examples superficially similar to Hudson’s yet different in structure and in distribution. The informal how many more years do you wanna go? and variants like it occur occasionally in the US, GB, AU, and KE. Variants of the standard How many more days and nights do you want to waste? also occur occasionally in the US, GB, BD, and ZA. As the results presented below reveal, the diffusion of want to/wanna in various grammatical patterns is often enough diverse throughout world Englishes. Whether this diversity fits a principled framework invites study as well.

Turning now to issues of substitutability and diffusion for want to/wanna in pragmatic contexts, particularly in indirect speech acts, analysis so far is incidental. Krug (2006: 147–148) reports hearing in California, “You wanna turn right … at the next corner,” and suggests that it is a command. Although this directive with wanna is common in the US, the corpus of Global Web-Based English (GloWbE) remarkably contains instances only of want to, as in AU You want to turn left at this intersection. A further issue, related to instances of want to and wanna, concerns their functioning as speech acts (direct or indirect) or as something else. One utterance, PK You want to turn around but you can’t move, seems possibly an indirect speech act yet in context is part of an extend commentary that depicts a character in fright. Since utterances in GloWbE beginning with you want to (6503) and you wanna (234) are numerous, they provide ample possibility for distinguishing their use as indirect or direct speech acts from other functions.

The analysis of want to/wanna that follows after an account of the data in GloWbE presents four sets of results: phonological, grammatical, semantic, and pragmatic. Each set begins with a framework designed to address the issues of substitutability and diffusion. The sets, as outlined here, comprise four sections of results:

Set 1. Phonologic patterns of want to/wanna
  Framework: pronunciation of {o} in want to and {-a} in wanna at the end of clauses and in utterance final positions
  Substitutability: Conditions influencing the use of a full or reduced vowel in to, the occurrence of wanna
  Diffusion in world Englishes: range of occurrences

Set 2. Grammatical patterns of want to/wanna
  Framework: want to/wanna in main and subordinate clauses
  Substitutability: restrictions on uses of want to/wanna in particular, grammatical environments
  Diffusion in world Englishes: occurrences of want to/wanna by grammatical pattern
Set 3. Semantic patterns of want to/wanna
Framework: patterns of grammatical and pragmatic features that condition the meanings of want to/wanna
Substitutability: contextual constraints on choice and meaning of want to or wanna
Diffusion in world Englishes: occurrence of meanings for want to/wanna in identical and similar contexts

Set 4. Pragmatic patterns of want to/wanna in indirect speech acts
Framework: locutionary and illocutionary uses
Substitutability: constraints on choice of wanna or want to for various uses and contexts
Diffusion in world Englishes: occurrence of diverse functions for want to/wanna

2. Method

2.1 Data appraisal

The data in the Corpus of Web-Based Global English (GloWbE) provide, with some insufficiencies, useful enough evidence for examining issues of substitutability and diffusion. The chief advantage of the data is that examples of you wanna are spontaneous. But this advantage has two shortcomings that condition the possibility of reliable and comprehensive results. One impediment to achieving reliable results comes from the unconventional spelling wanna. The online spelling want to does not provide in itself a guide to actual pronunciation.³

To evaluate this difficulty, two measures offer a perspective on the inaccuracy in the GloWbE data. The first measure is a chi-square test on the proportion of want to/wanna occurrences in GloWbE’s varieties. The second is Krug’s report on the incidence of want to/wanna in the spoken and written sections of the British National Corpus (BNC). The result of the first test is significant: the likelihood of a similar distribution for want to/wanna across the twenty varieties in GloWbE is less than .001. This hugely disparate incidence of the forms want to/wanna at best yields a sketchy outline of their substitutability and diffusion throughout world Englishes. Krug’s findings show enormous disparities, too. The written corpus, consisting of texts labelled as ‘imaginative,’ ‘arts,’ ‘leisure,’ ‘science,’ and ‘world

³. Krug (2000: 38) notes the possibility of error in transcribing instances of wanna in the spoken component of British National Corpus. Since the spoken component had many transcribers and a “limited number of cross-checks (20%),” transcriptions are open to inaccuracy.
affairs,’ is a medium likely attuned more to standard than informal spelling. Not surprisingly, Krug (2000: 154) reports that of 22,102 instances in the written corpus, 191 take the form wanna, and 21911 take the standard want to. In contrast, the spoken corpus contains want to 7,851 times, wanna 2,381. This salient difference has a proportion that resembles that of the standard and informal spellings for GB in GloWbE: 44,939 for want to, 629 for wanna. Although a chi-square test also yields a difference greater than .001, wanna in GloWbE for GB occurs more freely among internet users than among writers in the BNC.

These findings caution against the use of statistical analysis of want to/wanna in the GloWbE corpus. Instead, this analysis proceeds on a somewhat different assumption. If for some varieties of world English the GloWbE corpus contains, for example, a use of wanna appearing at least once, that evidence is enough to contribute to a statement on its diffusion. Such a statement is provisional, however, since she wanna in she gonna do what she wanna do in one US online entry hardly indicates its presence elsewhere in American English. Infrequent occurrences (1 to 3), then, of wanna or want to in diverse, linguistic environments are possibly false positives. The GloWbE corpus may also lead to possibly false negative inferences. For example, all varieties of world English, except for LK, KE, TZ, contain instances of I wanna in relative clauses such as I get the feeling that I wanna say goodbye (US). Since the GloWbE corpus does not report I wanna in the relative clauses of these three varieties, the reliability of such absence is questionable. Also questionable, of course, is the absence from no more than three varieties of wanna or want to in diverse linguistic environments of world Englishes.

2.2 Sociolinguistic appraisal

GloWbE’s data present considerable difficulty in learning the attributes of speakers and auditors who express and receive the want to/wanna locutions. Identifying nationality, gender, age, education, or work of persons of senders and recipients in all the GloWbE varieties is subject to intractable impediments. The absence of requisite profiles precludes statistically supported, convincing sociolinguistic analysis.

2.3 Want to/wanna and distinctions between standard and informal grammar

The labelling of want to as standard, wanna as informal raises questions regarding the status of forms and structures collocated with them. Utterances recorded in the GloWbE data often exhibit a mixture of standard and informal features as in these two instances:
(5) You want to get that radical new hair cut then do it. (IN)
(6) You wanna get with me, you got ta listen carefully. (GB)

The absence of the subordinator *if* in (5) and (6) reveals a fluent border between standard and informal usage. In each instance the first clause depends on a pattern of intonation unaccompanied by *if* that identifies it as syntactically subordinate—an informal practice. Like (5) and (6), utterances throughout the GloWbE corpus often enough have grammatical features in *want to/wanna* utterances that are informal. Whether a contrast between standard and informal features affects issues of substitutability and diffusion of *want to/wanna* requires analysis.4

In Sections 3 and 4 nine tables present an overview of phonological and grammatical features and collocations of *want to* and *wanna*. Each of these tables has three divisions: (1) utterance examples, (2) a frequency listing for each feature or collocation under analysis, (3) a listing for patterns of diffusion. Each frequency listing appears in the Appendix, identified by table and by feature or collocation for *want to* and *wanna*. The listings in the body of the text for patterns of diffusion throughout the tables provide the relative frequency for each variety of the world Englishes under analysis. The following key identifies the relative frequencies: + for at least four instances of *want to* or *wanna*; - for the absence of any instances; * for one to three instances (false positive); ** for the absence in no more than three among the twenty varieties of any feature or collocation (false negative).

3. Results: Phonological features of standard *want to*, informal *wanna*

Three issues concern the pronunciation of *want to* and *wanna*. The first involves effects on the pronunciation of *want to* and on the occurrence of *wanna* occurring at various places in clauses. Quirk et al. (1985: 909) note that before ellipsis in the middle or at the end of clauses the pronunciation of *to* is either /tu:/ or /tʊ/. As for *wanna*, Krug (2000: 156) finds that it occurs less frequently in certain patterns of ellipsis than *want to* or *want*. His examples are … *if she wants to, she can get in touch with you* and *You can wait outside if you want*. Whether *wanna* ever occurs at the end of a subordinate clause in non-final position is not clear. As for *wanna* in utterance final position, Krug’s data from British English imply a general

---

4. Hudson (2006: 605) offers the utterance *I want, to (‘wanna) be honest, a bigger piece of cake* as an example of a structure that permits the standard but not the informal verb. Although in some varieties *I wanna a bigger piece of cake* is unacceptable, Table 3(c) lists several that now instance the informal as taking a direct object. In these instances *wanna* is form of verb that does not presuppose *to*, manifested morphosyntactically in the final –a.
infrequency that calls for further analysis. The results below (Tables 1 through 9) on the occurrences of wanna and want to reveal patterns of diverse diffusion.

Table 1. The occurrence throughout world Englishes of want to and wanna in three structural positions

<table>
<thead>
<tr>
<th>Structural Position</th>
<th>Utterance examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial</strong></td>
<td></td>
</tr>
<tr>
<td>1a. y/n interrogative</td>
<td>Want to see what is new in the latest version? (GH)</td>
</tr>
<tr>
<td>1b.</td>
<td>Wanna see it? (NZ)</td>
</tr>
<tr>
<td><strong>Medial</strong></td>
<td></td>
</tr>
<tr>
<td>2a. at close of clause</td>
<td>Make notes, if you want to, of things to remember (PH)</td>
</tr>
<tr>
<td>2b.</td>
<td>Here are four questions I ask myself when I wanna, gotta, would love to but CAN’T (AU)</td>
</tr>
<tr>
<td><strong>Final</strong></td>
<td></td>
</tr>
<tr>
<td>3a. at end of clause</td>
<td>You can attend both classes if you want to. (TZ)</td>
</tr>
<tr>
<td>3b.</td>
<td>You go ahead and call AAA if you wanna. (US)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

<table>
<thead>
<tr>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

As a whole, Table 1 extends Krug’s findings for British English to varieties worldwide. From a phonological perspective, the y/n interrogative occurs worldwide, its intonation likely consistent. Subordinate clauses containing wanna, in medial position or at the end of utterances, are far less predictable. They have little frequency in any of the varieties listed and go unrecorded in about half.

The second issue focuses on variations in juncture and pause for forms of want to and wanna. Here syntax may affect pause and the pronunciation of {o} in wantto, as in if you want to, you can add all kinds of active adventures to your safari. The quality of the vowel – [u], [ʊ], or [ə] – very likely, together with pause length, varies from one enunciation to another. Moreover, since if-clauses alone do not influence pause and pronunciation, the results below reveal other syntactic patterns that also contribute to these phonological alternatives and their occurrences in world Englishes. Also, the likelihood of /a/ is greater than the two other vowels, if none of the words in the clause before medial pause and none beginning
the clause that follows draws primary stress. In Table 2(b) below the occurrence of /ə/ instead of /u/ or /ʊ/ stems from a lack of prominent accent, either in the initial clause or on the first words of the next clause. The accentual pattern on traditional diamond shaped scoring gives prominence to the object of the clause, not to its subject or verb form.

As for wanna, its occurrence both at the end of clauses and utterances is also unpredictable. The phonologic elements that possibly influence the choice of wanna in mid or final position of utterances are unclear. One example — *it ain’t easy, livin’ like ya wanna, and it’s so hard to find peace of mind* (IE) – manifests in the assonantal string in *ya wanna and a supporting, phonic pattern. To cite another example — *I-you know, I don’t wanna –* the string of nasals and vowels also form an alliterative and assonantal pattern. In this example, too, and others a strong stress on *don’t or do* may influence the choice of a following wanna (all varieties but HK, GH, KE). Table 2 includes data, as well, on the diffusion of subordinate clauses containing want to or wanna, as listed under (a), (b), and (c):

**Table 2.** At the close of clauses, the vowels for to in want to, and the occurrence of wanna

<table>
<thead>
<tr>
<th>Vowel quality</th>
<th>Utterance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. /to/: /u/ or /ʊ/ either vowel is possible</td>
<td>If you really want to, you can avoid falling in love (GH) If they want to, it’s play. If they don’t, it’s work. (MY) When they want to, even teenage boys can do it, (GB)</td>
</tr>
<tr>
<td>b. /to/: /ə/</td>
<td>'All right. If you want to,’ said the boy. (US) If you want to, you can do the traditional diamond worldwide shaped scoring, (IE) When you want to, you can use the Surface like a full laptop (CA)⁵</td>
</tr>
<tr>
<td>c. {wanna}/: /ə/</td>
<td>if you do wanna, then do it, who really cares. (AU) you can get it if you wanna, boy, the glock or the nine (PH)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

<table>
<thead>
<tr>
<th>Patterns of Diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
</tr>
<tr>
<td>a.</td>
</tr>
<tr>
<td>b.</td>
</tr>
<tr>
<td>c.</td>
</tr>
</tbody>
</table>

⁵. The occurrence of /ə/ in these examples is likely due to the influence on it of the accompanying phonologic or grammatic patterns.
The second example in Table 2(c) does not evidence the influence of assonance or nasalization, yet the pattern if you wanna, found no more than once in three varieties, is very likely a false positive.

The third issue in this section of results centers on possible phonological aspects of the spellings {wanto}, {wan to}, {wanna} and {wann}. In regard to {wan-na}, the GloWbE data offers these contrasting examples: You wanna piece of me? (AU) and What I wanna to do? (HK). The spelling {wanto} stands, before to, as a single unit in I really wanto to extend my help (US). This example suggests a substitution for a single, released [t]. This substitution differs from the two t’s in want to, the first an unreleased stop, the second aspirated, a slight juncture in between. Another variant of want to occurs in I don’t wan to speak of other atrocities: this {to} in its phonic form possibly has an aspirated [t]. All these variants are misspellings, but even so they possibly stem from the influence of pronunciation.6 Table 3 outlines for wanto, wan to, wanna, and wann both their unexpected grammatical uses and their diffusion in world Englishes.

Krug (2000: 211–214) argues cogently that the resemblance of wanna to gon-na and gotta is not mere happenstance, but owes much to phonological association, frequency, and grammatical structure. The phonological features of wanna (see Section 1) apply as well to gonna and gotta (the /t/ of gotta homorganic with /n/ in the other two words). All three forms recur with great frequency in world Englishes; all appear as modals before verbs in the infinitive.7 The instances of wann are less numerous (about a hundred in the GloWbE corpus), found in fewer than half the English varieties, but intriguingly have a phonological shape akin to the modals can, wont, and don’t. Some instances of wanna in Table 3(c), particularly before pronouns and nouns work functionally as alternates to the standard want, the – a retained simply as a phonological element. Whether wanna, as an informal, grammatical alternate to want becomes more frequent, currently remains unclear.

---

6. Although almost all languages include /t/ in their phonemic inventories, allophonic features differ. The notes on the allophones of /t/ in this paragraph are, of course, speculative, especially since the linguistic backgrounds of the internet users are unknown.

7. In the conclusion of his study Krug says, “Wanna, gonna and gotta have assumed various features that are typical of modal verbs in general, and of the category of emerging modals in particular” (2000: 252). His study presents an argument that several “quasi-modals,” such as be going to, have got to, have to, and want to “are assuming some, but not all, features that are typical of the core members of the English modal paradigm” (2000: 2).
Table 3. *wanto, wan to, wanna, and wann* in grammatical environments and in world Englishes

<table>
<thead>
<tr>
<th>Form</th>
<th>Utterance Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. <em>wanto</em></td>
<td></td>
</tr>
<tr>
<td>before the infinitive verb</td>
<td>they <em>wanto</em> support their families (PK)</td>
</tr>
<tr>
<td>before the marker <em>to</em></td>
<td>I <em>wouldnt</em> <em>wanto</em> to do it in Columbo (LK)</td>
</tr>
<tr>
<td>b. <em>wan to</em></td>
<td></td>
</tr>
<tr>
<td>before the infinitive verb clause final</td>
<td>We <em>wan to</em> talk about all of it. (PK)</td>
</tr>
<tr>
<td></td>
<td>u can stick with monster carnival straight to lvl50 if you <em>wan to</em>. (SG)</td>
</tr>
<tr>
<td>c. <em>wanna</em></td>
<td></td>
</tr>
<tr>
<td>-a as contraction of {<em>t to</em>}</td>
<td>i wanna c it n e way (JM)</td>
</tr>
<tr>
<td>-a reduplication</td>
<td>you wanna a see a alien (US)</td>
</tr>
<tr>
<td>-a as to + article</td>
<td>that why u <em>wanna</em> small place? (HK)</td>
</tr>
<tr>
<td></td>
<td>wanna observation on some general issues (ZA)</td>
</tr>
<tr>
<td>-a as phonological element</td>
<td>anyway just wanna this(JM)</td>
</tr>
<tr>
<td></td>
<td>I wanna a latte. (SG)</td>
</tr>
<tr>
<td></td>
<td>They wanna u be safe (PK)</td>
</tr>
<tr>
<td></td>
<td>i don’t <em>wanna</em> you to feel this pain (US)</td>
</tr>
<tr>
<td>d. <em>wann</em></td>
<td></td>
</tr>
<tr>
<td>as modal</td>
<td>Why you <em>wann</em> try something new (PH)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

| | US | CA | GB | IE | AU | NZ | IN | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
| (a) | + | − | * | * | * | − | − | − | − | * | − | − | − | − | − | − | − | − | − |
| (b) | + | + | + | * | + | + | * | + | + | + | + | * | + | + | + | * | * | + | + |
| (c) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| (d) | + | − | + | * | − | − | − | − | − | * | + | * | − | * | − | * | − | − | − | − |

4. Results: Grammatical features of standard *want to*, informal *wanna*

The results with respect to grammatical features of *want to* and *wanna* include categories of mood, voice, subordinate clauses as well as notes on collocation. Under these categories, too, the analysis presented examines issues of standard and informal use and the extent of diffusion of select grammatical features. The survey focuses on *want to* and *wanna*, not on any of the other variants found under phonology (3.1).
4.1 Categories of mood

These categories – imperative, indicative, and interrogative – reveal in Table 4 diverse patterns for want to and wanna, many substitutable, others not. Surprisingly, some of the patterns discussed in studies by Hudson and by Pullum occur sparingly or not at all in the GloWbE corpus. Among imperatives, Pullum’s example (1997: 81) – you wanna watch yourself – has a somewhat analogous example in the GloWbE data: After all you don’t wanna be judged yourself (GB), yourself, here, an emphatic reflexive. Among interrogatives both Hudson (2006: 608) and Pullum (1997: 90) have Who do you wanna see?, a pattern with wanna not found in the GloWbE corpus. Among indicatives, Pullum (1997: 91–92) discusses I wanna very much go to the game this evening as exemplifying in wanna a morpheme derived from want + to. The point of interest here is the intrusion of an intensive adverb between the verb wanna and its complement. The GloWbE corpus has two comparable examples of this structure: I wanna fucking go there (ZA) and I wanna just definitely do a program (US). For many speakers, Pullum finds structures that have want + intensive adverb + infinitive complement, as in I want to very much go to the game unacceptaible, yet GloWbE data show otherwise. Here are three examples, generally acceptable or not, of at least a dozen: I want to really support him (KE); I want to truly follow you (BD); I want to particularly welcome the relatives … (IE).

For want to and wanna Tables 4, 5, 6 list structures under the headings imperative, indicative, and interrogative, all accompanied by sample utterances and frequencies of their diffusion in world Englishes. Instances of want to/wanna as imperatives always occur with you as subject and function as directives or offers of advice: so you want to make sure that you have plenty of work to do (HK). In contrast, an utterance such as You want to hit me, he smiles. Come, here. (GB), exemplifies a desiderative attributed to the person addressed. Unlike directives and offer of advice, desideratives with you as subject are indicative in mood.

8. Pullum occasionally cites other linguists’ usage but his data are unattributed. Hudson (2006: 607) undertook an “informal e-mail survey” in determining variants “in the pronunciation of wanna and to” by sixty-two linguists and phoneticians. Otherwise, his data are his own and his colleagues’.

9. A desiderative differs from such utterances as You want to know what an arrestee’s condition is before they go into the van, for a number of reasons (NZ). The subject you is here impersonal, the utterance as a whole a statement on a regulatory procedure. Examples of these two patterns appear in the table for indicative utterances.
Table 4. Imperatives for *want to* and *want* in world Englishes

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Utterance Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. <em>want to</em></td>
<td></td>
</tr>
<tr>
<td>1 you want to INF + NP</td>
<td>You want to make the offer (PK)</td>
</tr>
<tr>
<td>2 you want (PN/NP) to</td>
<td>You want documentation to back yourself up so there will be no problems in the future (NG)</td>
</tr>
<tr>
<td>negated modal</td>
<td>You don’t want to send the message that it’s o.k to have sex at a young age (TZ)</td>
</tr>
<tr>
<td>b. <em>wanna</em></td>
<td></td>
</tr>
<tr>
<td>1 you wanna + INF + PN</td>
<td>You wanna do it with a great guy (US)</td>
</tr>
<tr>
<td>2 you wanna + INF + NP</td>
<td>You wanna go get that girl</td>
</tr>
<tr>
<td>negated modal</td>
<td>you don’t wanna get me started here (NG)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

<table>
<thead>
<tr>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

Table 5. Patterns of *want to* and *wanna* in world Englishes as indicatives of principal clauses

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Utterance Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. <em>want to</em></td>
<td></td>
</tr>
<tr>
<td>Pr/NP + want to</td>
<td>Utterance Example</td>
</tr>
<tr>
<td>1 I want to</td>
<td>I want to show you a video (US)</td>
</tr>
<tr>
<td>2 we want to</td>
<td>we want to help solve that problem (HK)</td>
</tr>
<tr>
<td>3 you want to INF + NP</td>
<td>You want to sugar coat every situation and say that diplomacy (leniency) is always the answer. (AU)</td>
</tr>
<tr>
<td>you want (PN/NP) to</td>
<td>You want me to ditch my original plan, fine (NG)</td>
</tr>
<tr>
<td>you want to INF +/- PN</td>
<td>You want to lock me up, fine. (GB)</td>
</tr>
<tr>
<td>negated modal</td>
<td>You don’t want to tell stories, fine (US)</td>
</tr>
<tr>
<td>4 he wants to</td>
<td>He wants to leave the control in safe hands (LK)</td>
</tr>
<tr>
<td>5 she wants to</td>
<td>she wants to run it in Japan (NZ)</td>
</tr>
<tr>
<td>6 he want to</td>
<td>He want to buy our Soft-150. (JM)</td>
</tr>
<tr>
<td>7 she want to</td>
<td>She want to try and get us to work again.(CA)</td>
</tr>
<tr>
<td>8 it want to</td>
<td>It want to get out of business (TZ)</td>
</tr>
<tr>
<td>9 they want to</td>
<td>They want to be remembered here tonight (CA)</td>
</tr>
</tbody>
</table>
The substitutability and diffusion of *want to* and *wanna* in world Englishes

10  NP + want to  
    Sister want to donate marrow to save his sister (JM)  
    People want to know about us. (US)

Patterns of Diffusion

|   | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 2 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 3 | +  | +  | +  | +  | +  | *  | +  | +  | +  | *  | +  | +  | +  | +  | *  | +  | +  | +  | *  |
| 4 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 5 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | *  | +  |
| 6 | +  | +  | *  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | *  | +  | +  | +  | *  | +  |
| 7 | +  | +  | *  | +  | −  | +  | *  | +  | *  | +  | +  | *  | +  | *  | +  | *  | *  | +  |
| 8 | *  | −  | +  | *  | *  | *  | *  | −  | −  | −  | +  | −  | −  | −  | +  | *  | *  | −  | −  |
| 9 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 10| +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |

a’. wanna

Pronoun + wanna  
Utterance Example

<table>
<thead>
<tr>
<th></th>
<th>I wanna</th>
<th>We wanna phone your mother***ers (ZA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I wanna</td>
<td>I wanna know their conversations (IE)</td>
</tr>
<tr>
<td>2</td>
<td>we wanna</td>
<td>We wanna phone your mother***ers (ZA)</td>
</tr>
<tr>
<td>3</td>
<td>you wanna</td>
<td>You wanna run or scream or cry, but something’s locking you up (PH)</td>
</tr>
<tr>
<td>4</td>
<td>he wanna</td>
<td>he wanna show J15 that he must pull up his socks (ZA)</td>
</tr>
<tr>
<td>5</td>
<td>she wanna</td>
<td>she wanna buy her own dress and shoes (AU)</td>
</tr>
<tr>
<td>6</td>
<td>it wanna</td>
<td>it wanna be more simple for members to use (BD)</td>
</tr>
<tr>
<td>7</td>
<td>they wanna</td>
<td>They wanna take me out for lunch (US)</td>
</tr>
<tr>
<td>8</td>
<td>NP + wanna</td>
<td>My co-worker wanna ask her why she doesn’t go home (CA)</td>
</tr>
<tr>
<td></td>
<td>People wanna get hung over one word, fine (JM)</td>
<td></td>
</tr>
</tbody>
</table>

Patterns of Diffusion

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>**</td>
<td>+</td>
<td>**</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>**</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>**</td>
</tr>
</tbody>
</table>

* ** want to **

People want to know about us. (US)
Krug (2000: 157–159) presents data (not conclusive) to show that regular modal auxiliaries precede such lexical verbs as *give*, *tell*, and *find* far more often than *want to* and *wanna* in the spoken BNC. Moreover, a fixed pattern governs the occurrence of regular modal auxiliary, *want to*/wanna, and lexical verb. He concludes that *want to*/wanna and lexical verbs comprise separate classes.

Double Modal10

b. *want to*

| Utterance Example | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|-------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 can want to they can want to conquer life’s challenges (JM) | + | + | + | * | * | * | * | * | * | * | * | ** | ** | + | + | * | | | |
| 2 could want to you could want to consider the elements (LK) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 3 may want to you may want to call a tow truck company (LK) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 4 might want to you might want to tweet about a URL (GH) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 5 shall want to I shall want to give you my own favourite (AU) | * | − | + | − | * | − | * | − | − | − | − | − | − | * | − | * | * | * | − |
| 6 should want to You should want to learn new things (BD) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 7 will want to he will want to win that tournament (PK) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 8 would want to I would want to consider the classification (TZ) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 9 ’ll want to You’ll want to weigh the pros and cons (PK) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 10 ’d want to he’d want to represent them (NG) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |

Patterns of Diffusion

b’. *wanna*

| Utterance Example | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|-------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 can wanna not found | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 2 could wanna not found | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 3 may wanna community member may wanna see those files (TZ) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 4 might wanna U might wanna use ur usual ones (GH) | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 5 shall wanna not found | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 6 should wanna not found | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |

10. Krug (2000: 157–159) presents data (not conclusive) to show that regular modal auxiliaries precede such lexical verbs as *give*, *tell*, and *find* far more often than *want to* and *wanna* in the spoken BNC. Moreover, a fixed pattern governs the occurrence of regular modal auxiliary, *want to*/wanna, and lexical verb. He concludes that *want to*/wanna and lexical verbs comprise separate classes.
7 will wanna  
8 would wanna  
9 ’ll wanna  
10 ’d wanna  

everyone will wanna be your friend (NZ)  
you would wanna play movies over the network (IN)  
he’ll wanna go over and say hi to her (NZ)  
I’d wanna see it (CA)  

Patterns of Diffusion

| VP + Ø … | I think I want to go (HK) |
| VP + that … want to | You know that you want to see it. (IE) |
| VP + wh …. want to | They know what they want to eat (HK) |
| Others will ask which you want to run it as – debit or credit. (JM) |
| I see which holes I want to fill with what. (CA) |
| Others know when they want to go. (IN) |
| you know where you want to operate (JM) |
| You choose who you want to see (GH) |
| we can see why you want to stay (HK) |
| you must understand how they want to be rewarded (IN) |

Patterns of Diffusion

| VP + Object Clause |
| VP + Ø … |
| VP + that … want to |
| VP + wh …. want to |
| Others will ask which you want to run it as – debit or credit. |
| I see which holes I want to fill with what. |
| Others know when they want to go. |
| you know where you want to operate |
| You choose who you want to see |
| we can see why you want to stay |
| you must understand how they want to be rewarded |

Patterns of Diffusion

| VP + Ø … | I think I want to go (HK) |
| VP + that … want to | You know that you want to see it. (IE) |
| VP + wh …. want to | They know what they want to eat (HK) |
| Others will ask which you want to run it as – debit or credit. (JM) |
| I see which holes I want to fill with what. (CA) |
| Others know when they want to go. (IN) |
| you know where you want to operate (JM) |
| You choose who you want to see (GH) |
| we can see why you want to stay (HK) |
| you must understand how they want to be rewarded (IN) |
c’. wanna 

Utterance Example

1. VP + Ø … You feel you wanna die (GB)
2. VP + that… wanna I just don’t know that i wanna wait (US)
3. VP + wh … wanna people are gonna say what they wanna say (US)
4. VP + which not found
   VP + which NP you can choose which path you wanna take (AU)
5. You know when you wanna continue reading (MY)
6. You know where you wanna go (GB)
7. I know who you wanna be (KE)
8. I don’t know why you wanna change that (ZA)
9. I dont care how you wanna pretty it up. (US)

Patterns of Diffusion

<table>
<thead>
<tr>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
</tr>
<tr>
<td>+</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>+</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

Copula + Predicate Complement

d. want to

Utterance Example

1. BE + that … want to The thing is that they want to break the resistance. (TZ)
2. BE + what This is what I want to do. (IE)
3. BE + when This is when you want to tell or advise the public (HK)
4. BE + where This is where I want to be. (IN)
5. BE + why that is why I want to help them (GH)
6. BE + how the bottom line is how you want to use it (KE)

Patterns of Diffusion

<table>
<thead>
<tr>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
The substitutability and diffusion of *want to* and *wanna* in world Englishes

Table 6. Patterns of mood for *want to* and *wanna* in world Englishes: Interrogative

<table>
<thead>
<tr>
<th></th>
<th><em>want to</em></th>
<th><strong>Utterance example</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y/n … want to</td>
<td>Do you want to have a great life? (ZA)</td>
</tr>
<tr>
<td>2</td>
<td>Don’t you want to see hear and feel nature? (JM)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Want to … ?</td>
<td>Want to take a hike? (GH)</td>
</tr>
<tr>
<td>4</td>
<td>You want to …?</td>
<td>You want to go off and do a graffiti book? (ZA)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

```
US  CA  GB  IE  AU  NZ  IN  LK  PK  BD  SG  MY  PH  HK  ZA  NG  GH  KE  TZ  JM
1  +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +
2  +   +   +   *   +   *   +   +   *   *   *   *   *   *   *   +   +   +   +   +
3  +   +   +   +   +   +   +   +   +   +   *   *   *   *   *   *   *   *   *   *   *
4  +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +
5  +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +
6  +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +
7  +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +   +
```

<table>
<thead>
<tr>
<th></th>
<th><em>wanna</em></th>
<th><strong>Utterance example</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y/n … wanna</td>
<td>Do you wanna be in the Olympics? (TZ)</td>
</tr>
</tbody>
</table>

Table 6. Patterns of diffusion for *want to* and *wanna* in world Englishes: Interrogative

```
US  CA  GB  IE  AU  NZ  IN  LK  PK  BD  SG  MY  PH  HK  ZA  NG  GH  KE  TZ  JM
1  *   −   −   −   −   −   −   −   −   −   *   −   −   *   −   +   −   −   −   −   −
2  +   *   +   +   *   +   +   *   −   −   −   *   *   *   *   *   +   −   −   −   −   −
3  −   −   −   −   −   −   −   −   −   −   −   *   −   −   −   −   −   −   −   −   −   −
4  +   −   *   *   *   *   *   −   *   −   −   −   *   −   −   −   +   −   −   −   −   −
5  +   *   −   *   −   −   −   −   −   −   −   *   −   +   *   −   −   −   −   −   −   −
6  *   −   *   −   *   −   −   −   *   −   −   *   −   −   *   −   −   *   −   *   −   −
7  −   *   −   *   −   −   −   −   −   −   −   −   −   −   −   −   −   −   −   −   −   −
```
2 Don’t you wanna check it? (US)
3 Wanna …? Wanna see some photos? (JM)
4 You wanna …? You wanna trash them? (BD)

Patterns of Diffusion

<table>
<thead>
<tr>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>−</td>
<td>*</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>*</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>*</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>*</td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>*</td>
</tr>
</tbody>
</table>

b. want to wh/…

Utterance example

1 who Who do you want to be sworn in? (NG)
2 which Which do you want to lose? (AU)
3 which + NP Which neighbourhood do you want to live in? (CA)
4 what What do you want to say? (PH)
5 when When do they want to be finished? (TZ)
6 where Where do you want to see yourself in 10 years' time? (BD)
7 why Why do you want to see me alone tomorrow? (JM)
8 how How do you want to spend it? (US)
9 how + many How many do you want to be carried away by? (GH and NG)
10 how + many + NP How many friends do you want to invite? (SG)
11 how + much How much do you want to raise, and why? (CA)
12 how + much + NP How much time do I want to commit? (KE)
The substitutability and diffusion of *want to* and *wanna* in world Englishes

b’. *wanna* wh/…

<table>
<thead>
<tr>
<th></th>
<th>Utterance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>who not found</td>
</tr>
<tr>
<td>2</td>
<td>which not found</td>
</tr>
<tr>
<td>3</td>
<td>which + N not found</td>
</tr>
<tr>
<td>4</td>
<td>what What do you wanna do? (IN)</td>
</tr>
<tr>
<td>5</td>
<td>what + pronoun What you wanna’ buy with that money? (MY)</td>
</tr>
<tr>
<td>6</td>
<td>when not found</td>
</tr>
<tr>
<td>7</td>
<td>when + pronoun not found</td>
</tr>
<tr>
<td>8</td>
<td>where Where do you wanna go? (only US)</td>
</tr>
<tr>
<td>9</td>
<td>where + pronoun Where you wanna do this? (only GB)</td>
</tr>
<tr>
<td>10</td>
<td>why why do you wanna be a waiter? (only US)</td>
</tr>
<tr>
<td>11</td>
<td>why + pronoun why you wan na control our lifes if you can’t provide a better in india? (IN)</td>
</tr>
<tr>
<td>12</td>
<td>how How do you wan na do that? (AU)</td>
</tr>
<tr>
<td>13</td>
<td>how + pronoun not found</td>
</tr>
<tr>
<td>14</td>
<td>how many not found</td>
</tr>
<tr>
<td>15</td>
<td>how many + NP not found</td>
</tr>
<tr>
<td>16</td>
<td>how much how much do you wan na do this? (only GB)</td>
</tr>
<tr>
<td>17</td>
<td>how much + NP not found</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

|   | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 4 | +  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | —  | —  | —  | —  |
| 5 | *  | —  | —  | *  | —  | *  | —  | *  | *  | *  | *  | *  | —  | —  | —  | *  | —  | —  | —  |
| 11| *  | —  | —  | —  | *  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  |
| 12| —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  |
| 16| *  | —  | —  | —  | *  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  | —  |

The result most salient in Tables 4, 5, 6 lies in the contrasts among structures of mood that instance *want to* and *wanna*. Especially in wh/interrogative structures (Table 6), the paucity of *wanna* is striking. For imperatives, too, *wanna* occurs in little more than half the varieties of English, negated or not. The examples followed by *fine* Table 6 (a and a’), are open to construal as versions of *If you want to lock me up, that’s fine* and *If you want to go crazy, that’s fine*. or perhaps as interrogatives. The decision to list these as indicatives is somewhat arbitrary, subject to the influence of context. The utterance *You want to go crazy, fine*, for example, directly

---

11. Krug (2000: 140–141) notes that in the spoken BNC a random sample of 500 instances does not yield an interrogative with clause fusion such as “Who do you want to succeed?”.
preceding *But don’t expect me to starve*, invites more than one grammatical analysis. This ambiguity also holds true for several other utterances listed under *want to* and *wanna* in Table 5.

### 4.2 Want to and wanna in relative and subordinate clauses

The diffusion of these structures in world Englishes, especially those with *wanna*, does not lend itself to reliable analysis. The incidence of false positives and false negatives diminishes the likelihood of findings representative of characteristic practices found in varieties of the GloWbE corpus. As with the array of false positives and negatives for structures included in the tables on mood, so in Table 7 on relative clauses and Table 8 on subordinate clauses, unpredictability is a prominent finding.

#### Table 7. Relative clauses for *want to* and *wanna* in world Englishes

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Utterance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. <em>want to</em></td>
<td></td>
</tr>
<tr>
<td>1 that</td>
<td>Yes as a physicist that I want to be (TZ)</td>
</tr>
<tr>
<td>2 Ø</td>
<td>I guess everyone uses a blog the way they want to. (LK)</td>
</tr>
<tr>
<td>3 when</td>
<td>we should set the time when we want to see it (BD)</td>
</tr>
<tr>
<td>4 where</td>
<td>you should first determine the measurement of the area where you want to set them. (HK)</td>
</tr>
<tr>
<td>5 which</td>
<td>We have a team goal which we want to achieve (IN)</td>
</tr>
<tr>
<td>6 who</td>
<td>Those are the people who I want to mix with (ZA)</td>
</tr>
<tr>
<td>7 why</td>
<td>There is a good argument why you want to be a bit more cautious (HK)</td>
</tr>
<tr>
<td>8 how</td>
<td>At the end is your choice how you want to do it. (JM)</td>
</tr>
</tbody>
</table>

#### Patterns of Diffusion

| | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
| 1 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 2 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 3 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 4 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 5 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 6 | + | + | + | + | + | + | + | * | + | * | + | * | + | + | * | + | * | + | + |
| 7 | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| 8 | + | * | + | * | + | − | * | − | * | + | * | − | − | * | − | − | − | * | − | − |
b. wanna

Utterance example

1. that I get the feeling that I wanna say goodbye (US)
2. Ø with whatever race you wanna, whatever gender you wanna, but wrong is wrong. (US)
3. when It’s just one of them days when I wanna be all alone (US only)
4. where This is the place where I wanna study (PH)
5. which on the top of the webpage which you wanna share (IE only)
6. who there are some people who I wanna buy food for (SG)
7. why Add to that a whole lot of other reasons why you wanna get up early (TZ only)
8. how it was about the principal how they wanna cheat us out of a few dollars (AU only)

Patterns of Diffusion

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Utterance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Subordinator + want to</td>
<td>Although I want to marry, I also like the single life (NG) Although you may want to give your children cherished memories, be careful not to push (US) It is like an addiction and difficult to leave, though I want to quit. (LK) Prices are reasonable, though you might want to haggle (TZ) I have no plans to go abroad because I want to find out about my husband (LK) That’s coz they want to feel biologically female. (MY) everyone should let them live how they want to. (BD)</td>
</tr>
</tbody>
</table>

The over-all result of Table 7 – that wanna hardly collocates with relative pronouns or adverbs other than that, supports the findings of the tables on mood. In contrast, the collocation of want to with these grammatical forms is far more widespread, if not uniformly worldwide. Table 8 explores this contrasting result as it applies to subordinators.

Table 8. Want to and wanna in the subordinate clauses of world Englishes

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Utterance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Subordinator + want to</td>
<td>Although I want to marry, I also like the single life (NG) Although you may want to give your children cherished memories, be careful not to push (US) It is like an addiction and difficult to leave, though I want to quit. (LK) Prices are reasonable, though you might want to haggle (TZ) I have no plans to go abroad because I want to find out about my husband (LK) That’s coz they want to feel biologically female. (MY) everyone should let them live how they want to. (BD)</td>
</tr>
</tbody>
</table>
6  if  If you want to go somewhere, you better wake up (US)
7  Ø  You want to dive you take the risks like I used to. (AU)
8  since  since I want to write about it, then I want it to be good. (JM)
9  unless  You never do anything unless you want to. (TZ)
10  until  But nobody can change until they want to. (LK)
11  till  Keep it till you want to have a happy marriage. (GB)
12  when  The time may come when you want to go back a bit (US)
13  where  it’s always going to lead you where you want to go (PK)
14  why  I’m not quite sure why you want to stop (GB)
15  adj + that cl.  Well, it’s nice that you want to do those things for me (JM)

Patterns of Diffusion

|   | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | *  | +  | +  | +  | +  | +  | +  | +  | +  |
| 2 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 3 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 4 | +  | +  | +  | +  | +  | +  | +  | *  | +  | +  | +  | *  | +  | +  | +  | *  | +  | +  | +  | +  |
| 5 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 6 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 7 | +  | *  | +  | *  | + * | *  | *  | *  | *  | +  | +  | *  | *  | *  | +  | +  | +  | +  | +  | *  |
| 8 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 9 | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 10| +  | +  | +  | +  | +  | +  | +  | *  | +  | +  | +  | *  | +  | +  | +  | +  | *  | +  | *  | −  |
| 11| *  | *  | +  | *  | −  | *  | −  | *  | −  | −  | −  | −  | −  | −  | −  | −  | −  | −  | −  | −  |
| 12| +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 13| +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 14| +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |
| 15| +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  | +  |

a’. Subordinator + wanna  Utterance example
1  although  I love your books although I wanna read more (GB)
2  though  Just set free the feline though you wanna hide her (JM)
3  because  I can’t do it because I wanna go shopping with my friend (AU)
4  cause (coz)  e mail me… coz i wanna read it! (JM)
5  how  its sweet how you wanna constantly be with me (IN)
6  if  If you wanna get a hat, get ahead (ZA)
7  Ø  People wanna get hung over one word, fine. (JM)
The substitutability and diffusion of *want to* and *wanna* in world Englishes

The co-occurrence of *want to* and *wanna* with subordinating clauses falls into a pattern similar to that of relative clauses, if somewhat less stark. As in Table 7, worldwide use of *want to*, not *wanna*, plainly characterizes most varieties of English. Yet instances of *wanna* with subordinators, as Table 7 indicates, seem more likely than with relative pronouns and adverbs (compare frequencies in Tables 7 and 8). Collocations of *wanna* with *which*, *how*, and *why*, as well as *although*, *though* and *until* appear infrequently. Overall, the diffusion of *wanna* in diverse grammatical structures and among diversities of English is a process still under way.  

12. In discussing grammatical functions of *want to* and *wanna*, Krug (2000: 141) finds that his data support the view that “changes start out in main clause, while subordinate clauses often retain conservative characteristics considerably longer.” The evidence found in GloWbE for *wanna* in subordinate clauses points to its increasingly diverse uses.
This process, an amorphous diffusion in the grammatical affinities of *wanna*, also has an unanticipated impact on substitutability in regard to *want to*. Potentially, among grammatical structures, even in third person singular patterns, *wanna* may occur wherever *want to* currently does. The instances in Tables 4 through 8 of *wanna* characterized as *not found* are hardly predictive. Nothing about the form, except the absence of grammatical inflection, structurally keeps it from possibly occurring in online transcripts wherever *want to* now does. And though, at least as a mnemonic, *wanna* counts as the informal counterpart of standard *want to*, the examples of utterances in the tables suggest that the distinction between informal and standard uses passes largely as a convenience.

4.3 Functional shift

As well as alternates in predicates, *want to* and *wanna* form fixed phrases before a following *be* to function as modifiers and nouns. Table 9 displays these functions, together with sample utterances and patterns of diffusion.

Table 9. Want to and wanna + be as modifiers and noun phrase

<table>
<thead>
<tr>
<th>Function</th>
<th>Modifier</th>
<th>Utterance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. <em>want to be</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 indefinite</td>
<td></td>
<td>a want to be farmer (NZ)</td>
</tr>
<tr>
<td>2 definite</td>
<td></td>
<td>the Want to Be Decorator (AU)</td>
</tr>
<tr>
<td>3 adverbial</td>
<td></td>
<td>for the want to be fair to everyone (IE)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion

|                | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|----------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|                | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  |
| 2              | *  |    | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  |
| 3              | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  | *  |

a’. *wanna be*  Utterance example

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 indefinite</td>
<td></td>
<td>a wanna be DJ (JM)</td>
</tr>
<tr>
<td>2 definite</td>
<td></td>
<td>the wanna be dad (US)</td>
</tr>
<tr>
<td>3 other determiners</td>
<td></td>
<td>this wanna be writer (PH)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>some wanna be Mayor (GB)</td>
</tr>
<tr>
<td>4 adverbial</td>
<td></td>
<td>the wanna be’ more efficiently designed’ (only US)</td>
</tr>
</tbody>
</table>

Patterns of Diffusion
The substitutability and diffusion of *want to* and *wanna* in world Englishes

Except for the uses *wanna be* as a modifier or indefinite noun (singular and plural), Table 9 suggests that functional shift appears mostly incipient among four or five varieties of English: US, GB, AU, PK. The fluidity of the data appears from the presence of *wanna be* as a noun but not as a modifier in ZA and KE.

5. **Results: The semantics of *want to* and *wanna***

In discussing the semantics of these standard and informal variants, Krug (2000: 147–151) draws for them six possible meanings: volition, intention, obligation, command, probability, and the hypothetical. The last two – probability and the hypothetical – unfortunately go without clear illustration. Although these meanings depend for Krug on issues of co-text and context (2000: 147), his analysis does not include a guide to his use of these terms. Since *want to* and *wanna* bear various meanings, ferreting out which applies in a given utterance benefits from an analysis of co-text. Fetzer’s brief overview (2012: 462–463) identifies co-text as...
“constructions embedded in adjacent linguistic constructions composing a whole clause, sentence, utterance, turn or text.” Further, “syntax is composed of structural units, for instance constituents and phrases […] – the linear ordering of the parts within a sequence which constitutes their grammatical function.” For the semantics of want to and wanna, co-text is the first consideration, often enough supplemented by reference to context. Krug notes that their “volitional modality” is currently their dominant sense. Utterances that depart from this sense of “volitional modality” require either parsing that accounts for any of the other senses for want to and wanna or pragmatic approaches (see Section 5).\footnote{13} Co-text, for example, helps to contrast the meanings “volitional” and “intentional” in these two utterances:

(7) It is so beautiful that you want to be there all the time. (AU)

(8) You’re doing the one thing that you want to do more than absolutely anything in the world. (GB)

Although the sense of want to in (7) may involve intention on the strength of the adjective phrase “so beautiful,” the infinitive “to be” indicates a desire possibly more a sentiment than something likely. In (8), the principal and noun clauses combine to underscore intention, the first identifying an ongoing activity, the comparative structure in the second affirming that the activity is fulfilling.

5.1 Want to and wanna in the sense of “intention”

Group 1 presents utterances that support, through features of co-text, a sense of intention for want to.\footnote{14} As a whole Group 1 provides a sample group of utterances, if not complete, for establishing linguistic features, some pertinent, others not, in determining a sense of intention for the meaning of want to. To begin with non-pertinent features, the mood of utterances, whether negated or not, need not contribute for want to a sense of intention. Besides the table’s utterances, mostly

\footnote{13. In particular, the discussion of Krug’s six senses for want to and wanna has this order: volition Section 5; intention (5.1); obligation (5.2); hypotheticals (5.3); probability (5.4). A discussion of command occurs in 6.}

\footnote{14. In the absence of frequencies, the term group replaces table in this section of semantics. As the discussion of utterances included in the section’s groups demonstrates, almost all the utterances included in a group require comment. To list frequencies with accompanying comment for each utterance requires an exhaustive presentation. The comments offered on the meanings of want to and wanna in the utterances listed do not readily augur overarching patterns applicable to frequencies.}
**Group 1. Want to as semantically “intention”**

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) choose which operating system you want to use every time you start your computer.</td>
<td>US</td>
</tr>
<tr>
<td>(b) There’s no question we want to go out there and blow everybody away with our seven-minute show.</td>
<td>CA</td>
</tr>
<tr>
<td>(c) You’re doing the one thing that you want to do more than absolutely anything in the world.</td>
<td>GB</td>
</tr>
<tr>
<td>(d) First run a text search on the specific word you want to test for co-occurrences.</td>
<td>IE</td>
</tr>
<tr>
<td>(e) For the time being we want to keep things running as per usual to avoid any loss of jobs.</td>
<td>AU</td>
</tr>
<tr>
<td>(f) customers load up a tray with the food they want to eat and then take the tray to a checkout.</td>
<td>NZ</td>
</tr>
<tr>
<td>(g) let’s say i want to search a letter start with’ C’, an then it will view all name or serial no that start with’ C’.</td>
<td>IN</td>
</tr>
<tr>
<td>(h) Each time you want to use the SDK you just need to add “require once”</td>
<td>LK</td>
</tr>
<tr>
<td>(i) Neha says, nothing is alright but I don’t want to talk right now (as Peehu is there)</td>
<td>PK</td>
</tr>
<tr>
<td>(j) Okay so we want to host this ebook you created online.</td>
<td>BD</td>
</tr>
<tr>
<td>(k) I want to know for sure they have the show before making my donation.</td>
<td>SG</td>
</tr>
<tr>
<td>(l) I want to warn everyone not to be deceived tonight.</td>
<td>MY</td>
</tr>
<tr>
<td>(m) But no way do I want to “Overstay” and run the risk of my family being booted out</td>
<td>HK</td>
</tr>
<tr>
<td>(n) of course I want to help and to support of my family from their financially needs</td>
<td>PH</td>
</tr>
<tr>
<td>(o) I want to make sure we get some decent bling for our bucks</td>
<td>ZA</td>
</tr>
<tr>
<td>(p) Obviously I want to settle down one day, marry and have kids.</td>
<td>NG</td>
</tr>
<tr>
<td>(q) Again, just in case we only live once, I want to make it worthwhile</td>
<td>GH</td>
</tr>
<tr>
<td>(r) Basically we want to make a car that drivers can feel safe to use.</td>
<td>KE</td>
</tr>
<tr>
<td>(s) I am sure if you have interest in this car you want to know the performance of it.</td>
<td>TZ</td>
</tr>
<tr>
<td>(t) Yes I want to look nice but I want people to see me and not only the dress.</td>
<td>JM</td>
</tr>
</tbody>
</table>

in the indicative, examples (a) and (d), in the imperative mood, express a general instruction directed at no one in particular.

As for the interrogative mood, one that also has want to with a sense of intention is this:

(9) Got something you absolutely want to get done, but keep dragging your feet about? (JM)
Moreover, neither affirmative nor negative forms in utterances particularly influence want to’s semantic value, as evident in (b), (j), (m), and (t) among others in the group. Further, want to is part of the verb phrase in eleven principal and nine dependent clauses in the group’s utterances, yet these, too, are not directly relevant to determining a sense of intention. Instead, a complex of grammatical and semantic features contribute to establishing co-texts (see the paragraphs immediately below) that support a meaning of intention for want to.

Unlike mood, clausal pattern, and negative/positive polarity, classes of infinitive, together with adverbs of conviction, sequence, time, or recurrent frequency convert want to’s usual sense of volition (according to Krug 2000: 147) into one of intention. The infinitive phrases directly following want to comprise five groups: 1. Task-related – (a), (b), (c), (d), (e), (f), (g), (h), (q) and (3); 2. Participatory – (i), (j), (k), (n), (p), (t, see); 3. Evaluative – (m, run the risk), (o), (q), (r), (s), (t, look nice); 4. Mental – (l), (s); 5. Stative – (m- overstay). Some groups overlap as in the instance of (7), participatory as well as mental.

Several adverbs occurring with want to in the sense of intention imply that fulfilling a task involves sequence, time or recurrent frequency. These often have a scope over entire utterances that include infinitives related to tasks projected: first in (d), for the time being in (e); each time in (h); tonight in (k); again in (q). Other adverbs of recurrent frequency or sequence occur as direct modifiers of task-related infinitives: every time in (a) and as per usual in (e); then in (f); right now in (i). Adverbs of conviction include absolutely in (c); for sure in (k); no way in (m). Wide scope adverbs of affirmation include okay in (j); of course in (n); obviously in (p); basically in (r); yes in (t). The opener there’s no question in (b) and the preterite phrase make sure in (o) work as adverbs of conviction do.

The utterance beginning let’s say (g) differs from the others in Group 1, because it posits an example of an intention expressed by want to. The sequence adverb then has a scope connecting the speaker’s supposed intention to a process that directly follows. Another instance of the same pattern is Let ‘s say you want to turn right (ZA). Here intention and process fuse together in the verb phrase. Similarly, the utterance beginning with I am sure (s) also differs from the others because it ascribes a conditional desire in the subordinate clause to the auditor before specifying a requisite intention. An analogous example, somewhat different in grammatical patterning is I am sure if you want to publish the pro life side you can contact the Pro life campaign (IE).

Although the typical subject pronouns in all these utterances are I and you, instances of we and they also occur. Apparently the sense of “intention” for want to does not depend on the pronominal subject. This survey falls short, since only one of the utterances (f) in Group 1 contains a noun as subject of want to. Another utterance, no wonder ur wife wants to divorce u (GH) indicates that nouns as subjects
of *want to* hardly exhaust the possibilities. As for *wanna* in the sense of intention, the utterances found in Group 1 exemplify verbs and adverbs somewhat different from those in Group 1.

**Group 2. Wanna as semantically “intention”**

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) this is something I absolutely REALLY wanna do</td>
<td>US</td>
</tr>
<tr>
<td>(b) But wait, first I wanna show you the game room.</td>
<td>CA</td>
</tr>
<tr>
<td>(c) Okay. Number one. Obviously I wanna ask you about Knight Rider. Hoff:</td>
<td>GB</td>
</tr>
<tr>
<td>(d) I wanna go for a wander around Dublin now – thanks lads!!</td>
<td>IE</td>
</tr>
<tr>
<td>(e) Select the people you wanna place in the first photo by holding the mouse left button and moving around the persons you wanna move. #</td>
<td>AU</td>
</tr>
<tr>
<td>(f) they have their own content platform apps and sure as hell don’t wanna start paying 30% to Microsoft for something that has been free for 20 years</td>
<td>NZ</td>
</tr>
<tr>
<td>(g) THIS time I wanna be sure and keep a track from the very beginning to avoid any such mishaps.</td>
<td>IN</td>
</tr>
<tr>
<td>(h) They say they wanna bring me in guilty For the killing of a deputy</td>
<td>LK</td>
</tr>
<tr>
<td>(i) So basically they wanna humiliate Iran and they put “we will welcome you with open arms” crap in there, to say that it is the sole fault of Iran</td>
<td>PK</td>
</tr>
<tr>
<td>(j) I wanna try the app and see if I can fake some emotions</td>
<td>BD</td>
</tr>
<tr>
<td>(k) Plus I’m flying off to TW on the 15th so of course I wanna look good</td>
<td>SG</td>
</tr>
<tr>
<td>(l) My sister is no longer here in this world. I wanna help my nephew to achieve his and my sister’s dream.</td>
<td>MY</td>
</tr>
<tr>
<td>(m) Wood said the band was on a grueling rehearsal schedule in Paris with the “nose to the grindstone.” “We wanna give 200 percent,”</td>
<td>HK</td>
</tr>
<tr>
<td>(n) I didn’t know these kind of bracelets exists but of course I wanna check it out and be one of those happy customer to testify positive outcome</td>
<td>PH</td>
</tr>
<tr>
<td>(o) I wanna fucking go there and see what is on the other side of the 2-inch step.</td>
<td>ZA</td>
</tr>
<tr>
<td>(p) Hi, i wanna have a beach party and I’m gonna be catering for 25 people more or less.</td>
<td>NG</td>
</tr>
<tr>
<td>(q) We obviously don’t wanna wear the same thing</td>
<td>GH</td>
</tr>
<tr>
<td>(r) I wanna delve into this scenario just a little bit, indulge me if you will</td>
<td>KE</td>
</tr>
<tr>
<td>(s) I wanna wait till the second hand reaches 60</td>
<td>TZ</td>
</tr>
<tr>
<td>(t) We wanna make sure everything is out of the way and that we don’t have any distractions.</td>
<td>JM</td>
</tr>
</tbody>
</table>

One aspect of the infinitive phrases listed in Group 1 that recurs in Group 2 is the incidence of pairs. The pairs in Group 1 include *go out/blow* (b), *load up/take* (f),
overstay/run (m), help/support (n), settle down/marry/have (p). Group 2 has be sure/keep (g), try/see (l), check out/be (n), go/see (o). These pairs contribute to a sense of intentionality. As for the classes of infinitives and adverbs in Group 2 they are much the same as in Group 1, with a few differences. The phrase go for a wander (d) has an adverb of purpose; in the first photo (e) is an adverb of place, fucking (o) an intensifier.

5.2 Want to and wanna in the sense of “obligation”

In citing instances of want to and wanna with the sense of obligation, Krug (2000: 148 and 276, n. 39) discusses

(10) We want to refine these categories. Existing taxonomies don’t suffice.

as also carrying a sense of necessity. The utterance’s subject we indicates, too, that the speaker of (10) accepts the obligation and necessity expressed by the predicate want to refine these categories. A further qualification for (10) is that its context suggests a requisite, procedural undertaking. Like (10), all varieties in GloWbE contain utterances with want to that bear a sense of obligation, that fit contexts associated with established procedures. In contrast, none of the varieties includes examples of wanna that involve a sense of obligation befitting established, regularized patterns. Instead, wanna and want to, also, as used in all varieties to indicate obligation, are both expressive of personal initiative. Further, the obligatory often enough includes intention, as in this utterance already cited: of course I want to help and to support of my family from their financially needs (PH). The adverbial of course, here, collocated with the double infinitives to help and to support, indicates a responsiveness to communal attitudes on obligation. Groups 3 and 4 present want to and wanna in the sense of obligation as associated with personal commitment.15 The lack of utterances that include wanna + inf applicable to institutional procedures is due less to any limits in linguistic co-text than contextual influences. Although wanna + analyze does not occur (see the TZ utterance with want to analyze), GloWbE has one instance of wanna + replace (note want to replace in the GH utterance above). Yet no obligation underlies I wanna replace them with flowy lacy curtains, but then realized why the original owners (this is a rented place) installed blinds instead of curtains (SG). This worldwide discrepancy in contexts associated with institutional practices attests to the non-substitutability of informal wanna for standard want to.

15. The words obligation and commitment often enough occupy a common semantic space. The OED offers as one possible definition of obligation the following: “An act or course of action to which a person is morally or legally bound; a duty, commitment”.

**Group 3.** *Want to* as semantically "obligation" in regularized, procedural contexts

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>President Bush, reacting to North Korea’s launch of several missiles, says, “We want to solve this problem diplomatically.”</td>
<td>US</td>
</tr>
<tr>
<td>caregiving responsibilities for parents of children with disabilities often continue much longer in life than for other families, and so certainly I want to recognize the need and the duties that carry on throughout a lifetime.</td>
<td>CA</td>
</tr>
<tr>
<td>Downside is now I want to refine it with the lessons learn’t and I see more elegant designs in the future.</td>
<td>GB</td>
</tr>
<tr>
<td>We urgently want to refine that research so that it may help to decrease suffering from human cancer</td>
<td>IE</td>
</tr>
<tr>
<td>We decided that we want to refine what we were doing, but also try to bring something new</td>
<td>AU</td>
</tr>
<tr>
<td>As always, such events are only possible with support from organisations and I want to acknowledge their commitment to sharing our vision</td>
<td>NZ</td>
</tr>
<tr>
<td>And while that is happening, I want to proceed with operation PIVOT-STRIKE as we planned.</td>
<td>IN</td>
</tr>
<tr>
<td>Even the large chain supermarket are the same. # But no one wants to accept this fact and change the attitude of their staffs.</td>
<td>LK</td>
</tr>
<tr>
<td>We also want to determine the type of action that can most effectively be initiated against it</td>
<td>PK</td>
</tr>
<tr>
<td>AQM’s quality platforms are projects implemented for customers who want to inspect all their orders outside factories.</td>
<td>BD</td>
</tr>
<tr>
<td>Then we want to compute the cdf of Y</td>
<td>SG</td>
</tr>
<tr>
<td>I want to measure the decibles of his shout.</td>
<td>MY</td>
</tr>
<tr>
<td>I want to analyze how these things affect the mindsets that allow for discrimination to occur.</td>
<td>PH</td>
</tr>
<tr>
<td>We’re a big company and we want to rebuild a reputation for innovation.</td>
<td>ZA</td>
</tr>
<tr>
<td>There are two schools of thoughts I want to examine in relation to given aid to Africa They want to replace the current multi-stakeholder approach, which supports the free flow of information in a global network.</td>
<td>NG</td>
</tr>
<tr>
<td>yes I said fat because it is one of the side effects and not only affects our esteem but there is heart issues and so many other things, we want to attend meetings and not be asleep all day The scholars doing research on Twitter and gender… want to analyze those same “contextual signals” for a deeper understanding of gender differences in language Here is my cleaning and maintenance procedure I use when I want to prepare and engine for a final at a major race</td>
<td>KE</td>
</tr>
<tr>
<td></td>
<td>TZ</td>
</tr>
<tr>
<td></td>
<td>JM</td>
</tr>
</tbody>
</table>

---

16. The sense of negative obligation in this use of *want to* parallels the force in another LK utterance: “Those who offended will genuinely be obliged not to repeat what they did to harm other parties”.
A sense of obligation hardly depends on established rules and practices; throughout the English speaking world personal, obligatory impulses and acts recur, expressed often enough by the alternates want to and wanna. Group 4, listing utterances that exemplify want to and wanna in the twenty varieties found in GloWbE, provides a sample of personal obligation. Often enough these alternates collocate with the same infinitive.

Group 4. Want to and wanna as semantically “obligations” in contexts of personal commitment

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a.1) I want to help the other kids who are worse off than I am</td>
<td>US</td>
</tr>
<tr>
<td>(a.2) I wanna help volunteer with recovery if anybody can give me and link or updates</td>
<td></td>
</tr>
<tr>
<td>(b.1) I want to teach my kids how important it is to help others.</td>
<td>CA</td>
</tr>
<tr>
<td>(b.2) I wanna stress out the enormous importance of discipline and hard work.</td>
<td></td>
</tr>
<tr>
<td>(c.1) I want to give my kids all I can.</td>
<td>GB</td>
</tr>
<tr>
<td>(c.2) I wanna pay full price because they work hard and I’m proud that they tried</td>
<td></td>
</tr>
<tr>
<td>(d.1) I want to give all young people and children and families and elders and disable people a better future in life</td>
<td>IE</td>
</tr>
<tr>
<td>(d.2) I just wanna clarify one thing. I am not pro-choice in the sense you are speaking</td>
<td></td>
</tr>
<tr>
<td>(e.1) I want to make sure my kids have a roof over their head and food on the table</td>
<td>AU</td>
</tr>
<tr>
<td>(e.2) I don’t deserve her, but I wanna work hard and show her that she does matter</td>
<td></td>
</tr>
<tr>
<td>(f.1) I want to personally fund helping the poor that we do through our work.</td>
<td>NZ</td>
</tr>
<tr>
<td>(f.2) I’m really serious about it (hunger relief). So far I’ve been able to feed 2.5 million (kids), but I wanna feed a billion.</td>
<td></td>
</tr>
<tr>
<td>(g.1) In the midst of all this I want to remember to teach my kids to trust their choices and their instincts.</td>
<td>IN</td>
</tr>
<tr>
<td>(g.2) I wanna serve the nation, but what if politicians don’t allow me to do so</td>
<td></td>
</tr>
<tr>
<td>(h.1) I want to do something for the poor children who are being deprived of their basic rights</td>
<td>LK</td>
</tr>
<tr>
<td>(h.2) It is not only Rajapakses but many others wanna do good for SL [Sri Lanka]. And they become crossed when misused taht badly</td>
<td></td>
</tr>
<tr>
<td>(i.1) I want to raise my kids as good Pakistanis, the kids who can work community development</td>
<td>PK</td>
</tr>
<tr>
<td>(i.2) all I can do is to give them whatever food that I have on my hands, though, I wanna give more, that time I was also a struggling, working single mother</td>
<td></td>
</tr>
<tr>
<td>Utterance</td>
<td>Variety</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>(j.1) I want to help the poor as you did.</td>
<td>BD</td>
</tr>
<tr>
<td>(j.2) Just wanna say this is not meant to offend anyone,</td>
<td></td>
</tr>
<tr>
<td>(k.1) I want to be there for my kids and i want to do something i love.</td>
<td>SG</td>
</tr>
<tr>
<td>(k.2) What I wanna try to bring across is that… It’s not okay to actually like 2 person at the same time!</td>
<td></td>
</tr>
<tr>
<td>(l.1) Do I want to support these girls? Of course.</td>
<td>MY</td>
</tr>
<tr>
<td>(l.2) For Allah. I wanna be a good mom.</td>
<td></td>
</tr>
<tr>
<td>(m.1) seeing it wrote down in a food diary will soon make you conscious that you want to adjust your consuming habits.</td>
<td>HK</td>
</tr>
<tr>
<td>(m.2) they allow her to go home and he or she wanna stay home to take care kids</td>
<td></td>
</tr>
<tr>
<td>(n.1) Someday I want to help PAL kids like me with the same dream.</td>
<td>PH</td>
</tr>
<tr>
<td>(n.2) Everybody wants to earn more, everybody wants to be successful, but they wanna do a good job</td>
<td></td>
</tr>
<tr>
<td>(o.1) I went home sometimes because I want to work for my kids.</td>
<td>ZA</td>
</tr>
<tr>
<td>(o.2) i really wanna get this, for my parents… i believe they’ll love this after all they need this</td>
<td></td>
</tr>
<tr>
<td>(p.1) I’m a mom. I want to take care of my kids and protect them</td>
<td>NG</td>
</tr>
<tr>
<td>(p.2) Me and my son is mellow. I’m his father, so I wanna show him the proper way because he looks up to me</td>
<td></td>
</tr>
<tr>
<td>(q.1) I want to help the less-privileged children in acquiring knowledge</td>
<td>GH</td>
</tr>
<tr>
<td>(q.2) I wanna give everything back to my mum, she’s sacrificed a lot for me</td>
<td></td>
</tr>
<tr>
<td>(r.1) I want to provide children with a safe place when their parents are in the field working</td>
<td>KE</td>
</tr>
<tr>
<td>(r.2) As i start looking forth to help you innovative improvements and will eventually show it website …# Only wanna state that this is very useful,</td>
<td></td>
</tr>
<tr>
<td>(s.1) Kenya is my home. I grew up there. I want to advise people to vote in peace.</td>
<td>TZ</td>
</tr>
<tr>
<td>(s.2) I feel compelled I wanna help other kids break the cycle.</td>
<td></td>
</tr>
<tr>
<td>(t.1) I want to help my mother and her six children.</td>
<td>JM</td>
</tr>
<tr>
<td>(t.2) It’s about me and my kids and I wanna save some kids’ lives in the process</td>
<td></td>
</tr>
</tbody>
</table>

Typically the pronominal subject in the Group 4 utterances, both for predicates with *want to* and *wanna*, is I. The occasional third person subject, as in (h.2) and (m.2), comes with clauses of co-text that provide useful, supplementary detail. For hypothetical meanings associated with *want to* and *wanna*, the pronominal choices, as discussed below in Section 5.3, are predominantly indefinite.
5.3 *Want to* and *wanna* in the sense of the “hypothetical”

Indefinite pronouns as subjects of *want to* and *wanna* appear in utterances that entertain hypothetical circumstances that preclude a clear sense of volition, intention, or obligation. A sense of possible obligation associated with *want to* appears in all the varieties, but not with *wanna*. One example,

(11) No one wants to trivialize or make light of seniors who can no longer walk without pain or care for themselves. (CA)

speaks to a common sense of obligation. As already indicated in Group 3, *want to* with a sense of obligation generally involves institutional commitment: with indefinite pronoun subjects, such instances occur in these sample varieties:

(12) If someone wants to be virtuous and be humane, and be a cityzen of the world, want equality among people… then there is the TEACHING OF THE LORD BUDDHA …! Be a real BUDDHIST…! (JM)

(13) Our civil right has been enshrined in the law. If anybody wants to take that right, it should be by the law (GH)

(14) Ergo, if anyone wants to be remembered as remarkably good, he must at all cost follow the rules no matter how hard it is to be good in the midst of temptations (PH)

Ten varieties also collocate an indefinite pronoun as subject of *wanna*, the verb in each instance semantically volitional. Among the nine, evidence for several indefinite pronouns occurs, the last two negatively volitional:

(15) If anyone wanna network, that would be great! (HK)

(16) But if someone wanna show the other aspect he is blamed of “rhetoric and undue emotionalism” (PK)

(17) I would like to share if anybody wanna try it. (IN)

(18) it is definite that if somebody wanna hire ppl having proven high IQ would turn towards local grads. (LK)

(19) If nobody wanna buy our adapters, we might just rent them out. (MY)

(20) IF no one wanna support you, I support you. (SG)

Another indefinite form is *people* as subject of *wanna*, evident in six varieties, the first expressive of volition, the other intent:
Group 5. “People” as subject of wanna in if utterances of volition or intent

**volition**

Utterance example

the comment space is getting really narrow... #If people wanna go the OpenFlow route, why not just build an API for routing protocols (US)

If people wanna hear what Norwegian music sounds like, listen to the folk music (GB)

Battletag is MossySloth and battle code is 2772 if people wanna add. (ZA)

**intent**

If people wanna take the chance, thats their risk. (LK)

But if people wanna buy, it is their biz (SG) really if people wanna do it and kill em selfs props to them. (JM)

A Jamaican English utterance without the subordinator if ambiguously borders the hypothetic and the presupposed:

(21) People wanna get hung over one word, fine

The ambiguity in (21) arises, owing to the concluding “fine,” a normative term that possibly derives from the speaker’s awareness of unhappiness with a word that for some has been inappropriately chosen. Inasmuch as personal pronouns in if clauses with want to and wanna occasion a similar ambiguity, utterances in 5.3 with subjects, either indefinite or generic, best illustrate a clear, hypothetical meaning.

5.4 Want to associated with probability

As with other senses, except “volitional,” a meaning of probability ascribable to want to largely depends on structural features in its co-text. The chief collocates grammatically linked to it are likely, possibly, and probably, as in this utterance:

(22) Criticism of her partner is only likely to make her want to defend him or her. (AU)

Nothing comparable with wanna appears in the GloWbE data. The scope of likely in (22) covers the entire predicate; in (23), want to is part of an extra-posed subject:

(23) it is most likely that you want to tell your prospective employer about yourself. (NG)

As well as carrying a sense of probability triggered by a grammatical bond between likely and want to, (23) also works often enough as an indirect speech act (see Section 6). Not all utterances that have a subject you bound together with a complement of probability are deontic. For example, (24) works primarily as an opinion:

(24) Possibly you want to survive a daydream that never feels happy (GH)
Whereas tell in (23) contributes to the force of an indirect speech act, survive in (24) does not. The diffusion of want to in the sense of probability is worldwide.

A second pattern that contributes to a sense of probability is the double modal predicate, as in

(25) If it were me, I’d want to wait at least a year between Puppies (ZA)

Here the subordinate clause contributes to a sense of probability generated by the double modal. All varieties of English have utterances like (25); few (AU, LK, SG, PH) have instances with wanna, as in

(26) If you are a musician you’d wanna stop there to get connected with localmusicians. (PH)

Utterance (26), like (23), has deontic force, although both appear in somewhat different grammatical structures. This substitutability involving semantics, however, has evident limits. As with uses of want to in the sense of meaning “intention” in regard to regularized, procedural activities, a like possibility for wanna is not demonstrable. Yet the few examples of wanna in the sense of probability suggest that the workings of diffusion in the varieties of worldwide English are still on-going.

The upshot is that as in the instances of grammatical structure, so in those concerning semantics, variation as a lively element in English speech everywhere remains potentially evident. Section 6 on the pragmatics of want to and wanna provides a further opportunity to examine the vibrancy of variation.

5.5 Want to and wanna in figurative use

Krug suggests (2000: 150) that the utterances Coolers? They wanna be on one of the top shelves somewhere incipiently embodies grammaticalization, since the subject pronoun has an inanimate rather than human reference. Yet the GloBwE corpus yields several examples of non-human subjects for want to or wanna that support a figurative construal. These include:

(27) For some reason, my hot water heats when it wants to. (US)

(28) It’s a simple waist strap with four suspended straps. The idea is the user controls how low they wanna sag (TZ)

Whether these and the few other instances in the corpus are false positives of a semantic (or grammatical) development in the varieties of English obviously awaits further sampling. In (27) and (28) the occurrence respectively of my and user implies a transference of human volition to the non-human, short of the
grammaticalization in Krug’s example. A further complication is the view given to *want to/wanna* as predicates related to animate, non-human life as in

(29) as slow as that of a colour-loving creature called chameleon, which always seems to be undecided on whether it wants to move forward (TZ)

The question here on the status of *it*, as a sign of grammaticalization, is probably open to dispute.

6. **On the pragmatics of *want to/wanna***

Verplaetse’s account presents a trenchant analysis of how *want to* and *wanna* work in speech acts. Relying on data in the BNC (Spoken Demographic), she argues that in illocutionary utterances, both the standard and informal predicates help to offer advice. The advice given, further, may indicate both the speaker’s perspective and an auditor’s. In some locutions, both perspectives are evident; Verplaetse (2010: 42) cites this example:

(30) well I know, well you want to label it so you know what’s on it … just reminding you, you say, oh you’ll say you didn’t remind me to write it down.

Other utterances offer “a *vicarious* statement of volition,” the speaker projecting advice simply for the benefit of auditors (2010: 41–42):

(31) San Francisco passengers, you wanna board this train

Here, the speaker, though not taking the train, conceives what auditors are seeking to know, not what is personally advisable. As utterances, (30) and (31) are open, short of richer, more fully explicit contexts, to differences of interpretation on the senses of *want to* and *wanna*. How much weight to give to the semantics of volition and the force of advice in these two utterances, without further context, remains debatable.

(32) And, like me, you care. You want to make a difference – for yourself, for your child, for someone you love, for someone who depends on you. And you can make a difference, you will, if you keep this cornerstone of

---

17 Verplaetse (2010: 38) cites as an internet example the utterance *I wanna you help me…* This instance of *wanna* before a pronoun and infinitive also occurs in three GloWbE US utterances, possibly a false positive. In a footnote, she cites two further examples, both taken from lyrics. In this study, utterances in lyrics do not count, because a choice of metrical and stylistic expression often enough affects their linguistic patterns. She also cites from BNC *you wanna’ve seen your face*, not found, however in the GloWbE corpus.
Autism Awareness in mind at all times. Ready? Here it is: # Behavior is communication. That’s it. That’s all. That’s everything. (US)

This sense of a shared purpose here presupposes that the speaker and auditor’s sense of want to is a volition held in common. As expressed, however, (32) leaves open the possibility that the speaker expects the auditor to undertake a joint or separate enterprise. A shared enterprise, obviously parallel, occasionally encompasses at least several people, as in

(33) It’s frustrating if like us, you want to get pregnant (sic) again straight away! (NZ)

The if clause anticipates the auditor’s concurrence in wanting to overcome an apparent obstacle. The like me, like us phrase has a limited currency with want to in ten varieties of English (also CA, GB, IE, AU, LK, SG, PH, NG, TZ), but none with wanna.

Verplaetse’s accomplished analysis spurs, moreover, inquiry into at least three unexamined issues on want to and wanna in pragmatic exchanges between speakers and auditors. The first concerns the relations discussed above in giving advice or inviting joint enterprise. The question to raise is whether the conditions underlying (30), (31), and (32) occur in all varieties of the GloWbE data, and if not, why not. The second centers on the occurrences, in respect to (30), (31), and (32), of want to and wanna. The third is to arrive at an overall view of want to and wanna in their pragmatic functions throughout the worldwide varieties of English.

6.1 Uses of want to/wanna and the speaker’s projections

As a deontic utterance, (31) expresses a speaker’s advice, those on the platform free to take it or not. In some contexts, however, the speaker’s utterance is not deontic but rather an empathetic comment on what the auditor desires, as in

(34) you’re tired and you’ve been pushing weights with Jeanette screaming at you, pushing through further than you ever thought you could go – you just wanna go home and sleep! (GB)

The semantics of wanna here, followed by two infinitives, resembles those found in Groups 1 and 2 as indicators of intention.18 The entire segment between dash and exclamation as a comment embodies an articulated presupposition. The speaker

18. In some instances, two apparent infinitives do not reliably help to contrast deontic utterances from comment: for example, You wanna try and keep em as far away from water as possible really (NZ). Here try and keep are not separate infinitives, but an example of pseudo-coordination (Quirk et al. 1985: 978–979) that is part of a deontic utterance.
voices the auditor’s unspoken desire. In the GloWbE corpus, the choice of wanna for comment or advisory purposes finds a limited range of analogs in the worldwide varieties of English. Four varieties (US, IE, ZA, GH) have tokens, both as advice and comment; six (CA, NZ, LK, PK, MY, NG) instance advice; seven (GB, AU, IN, SG, PH, KE, TZ) instance comment.

The standard want to combines with infinitives in deontic utterances worldwide, as in

(35) You want to give a voice to the new generation of mobile OS. Only possible way that this cause will be a success when you spread the word. A Nexus 4 will be instrumental means to engage and share and this contest is a good way to involve people. (IN)

Articulated presuppositions, also with want to, find expression worldwide, too, as in

(36) You want to kill me, but I don’t wish to kill you. (SG)

Unlike (34), the auditor’s presupposition in (36) that the speaker articulates meets with dissent. This voicing of exception also recurs throughout all the varieties of English in diverse ways. An example of an articulated presupposition that the speaker accommodates affably is as follows:

(37) You want to invite me for Chrismukkah but you don’t know how, so I will save you the trouble. (IE)

Whether the diverse ways of qualifying articulated presupposition lend themselves to systematic analysis awaits study.

6.2 From experience or counsel to the advisory use of want to and wanna

Utterance (30) includes you want to for the purpose of offering advice based on the speaker’s experience and on a concern that the auditor, if neglectful, will complain. Although the auditor’s possible complaint is not an uncommon response, the speaker’s advisory use of want to exemplifies a worldwide practice in the varieties of English. Often enough, as in (30), the advice offered stems from a personal experience or view, as in

(38) I feel big, because my atoms came from those stars. There’s a level of connectivity. That’s really what you want in life, you want to feel connected (JM)
Then you want to play in the Premier League, then the next step is to play for one of the top teams and then you want to play for your country. Thankfully I have done that – (GB)

Instead of a speaker’s I or me, the advice offered issues from a collective we or us:

Obviously those ‘Wedding Pictures’ are the trophies of life. And you want to look purrfect! No surprizes then that for most of us, weddings are the perfect trigger for wanting to get into shape. (BD)

Therefore practice may not make perfect, but it will make you persuasive – You want to rehearse out loud! We all sound fluid in our heads. (KE)

Comparable utterances that include you wanna are hardly evident. A single example, and so a false positive, occurs in Irish English (IE):

Those kind of people make me sick (bit of an exaggeration there, I know) wanna spare a thought for the person you are dealing with.

The absence of further examples expressed with you wanna is not altogether surprising, since half the varieties of English do not include them in utterances used to offer advice or a recommendation (see 5.1). As for advice or recommendations that suggest a shared undertaking or purpose, as illustrated in (32) and (33), ten varieties in all collocate the phrase like me or us and want to: US, CA, GB, IE, AU, LK, SG, PH, NG, TZ, JM. None collocates this phrase with wanna. As in grammar, and semantics, so in pragmatics, the diffusion of want to is more fully evident than wanna. This contrast, too, limits the substitutability of wanna for want to.

7. Discussion

The sections on results invite a review on the reliability of data found in GloWbE and a statement on their possible, theoretical value, especially in the light of Croft’s drawing from arguments on evolution (2006). The challenge presented by the data – in written form it may mislead on actual uses of want to and wanna – subsides partially because the patterns of diffusion for wanna reveal some consistency. Almost every table except 2 and every group except 3, despite a likely underrepresentation, contain instances of wanna diffused worldwide or in a substantial number of varieties. The unresolved challenge lies in determining whether wanna, found in more than three varieties (thus not a false positive), actually occurs in others. Although the instances of underreporting encourage further search for data, the over-all diffusion of want to and wanna suggests inferences that are plausible if not convincing.
The chief focus all through this study centers on the substitutability of wanna for want to throughout the varieties of world English. From the phonological, grammatical, semantic, and pragmatic incidence of these two alternatives, the salient finding is that they are substitutable, though uses of wanna are very likely greater than the data show, in almost every context. The absence of substitutability, as seen particularly in Group 3, is not, however, a minor exception. The universal preference for want to in the sense “obligation” in circumstances calling for regularized, procedural activities suggests that want to at times conveys a connotative propriety not associated with wanna. Yet this connotative propriety has its own constraint, inasmuch as wanna bears a sense of “obligation” in contexts that are social and familial (see Group 4). This difference in the application of propriety – wanna appropriate as a lexeme of obligation in respect to family and community but not to structured undertakings – is likely due to a tacit, not conscious, power of imposed behavior. Often enough, such power to command choices between comparable lexemes in familial and social circumstances occurs idiosyncratically. Oddly, the instance of wanna, in Irish English (42), recommending a shared approach, seems counterintuitive; maybe encouraging someone, even casually, to take such advice prompts the use of want to.

The second example of unlikely substitutability, as seen in Table 2, indicates few instances in utterances of wanna appearing at the end of a clause in a medial position. The GloWbE data yield five utterances in contrast to an extensive diffusion of want to. If the argument given – the unlikelihood of /ə/ occurring in medial position before a pause – requires instrumental testing, the discrepancy between occurrences of want to and wanna is unmistakable. Since wanna occurs worldwide in if clauses (see Table 8), an explanation that does not include phonologic constraints is hard to imagine.

A further inference on substitutability concerns uses of wanna exclusively, without any comparable example of want to. Table 3 contains such uses, the –a of wanna a wholly phonologic element in four environments, two of them found in more than three varieties of English. Although this finding seems almost incidental, it poses a challenge to analyses of the linguistic contrasts between want to and wanna that Hudson (2006: 604) presents. In his view, the –a of wanna regularly has the value to, yet the evidence of Table 3 suggests that this claim falls short. Apparently, some varieties, including American and British English (the two that supply the data for Hudson’s analysis), have developed innovations that his argument does not address. Yet the development of innovations in language is a recurrence that enters theoretic proposals, as Croft presents them.

Conceptually, innovation is for Croft (2006: 124) a development in the workings of linguistic form as it relates to its meaning and its grammatical properties; on the level of phonology, innovation emerges from the interplay between form...
and its phonetic properties. In regard to meaning and grammatical properties, wanna nowhere exhibits features that innovatively distinguish it from want to.

None of the tables on grammar and meaning or any utterance on pragmatic functions reveals wanna as a form innovatively distinct. On a phonological level, however, the phonetic realization of wanna (also wann) as unit forms, see Table 3(c), uninflected and unrelated to the infinitive marker to, is a fairly recent development. Between them, wanna and wann appear in fourteen varieties of English, although their frequencies online are infrequent, seven highest for wann in Jamaican English, ten for wanna in American English. Clearly, this range of distribution and infrequent occurrence provides little guidance to future instances of these two forms.

Croft (2006: 124) distinguishes innovative developments in form from propagation, the spread of forms in diverse communities. In regard to want to and wanna, the tables also attest to clear patterns of diffusion. In far more than half the entries, the tables show that want to occurs in worldwide communities, its uses diverse. For wanna, the patterns of diffusion vary considerably, sometimes not found, sometimes in use in all the varieties. The upshot is that the propagation of wanna is very likely still an ongoing phenomenon, its uses, if the tables on semantics are indicative, most common in informal circumstances.

If as Croft argues, innovative developments have strong, formal underpinnings, but propagation depends on social circumstances, then the patterns exhibited by want to and wanna provide an intriguing perspective on theory. This perspective suggests that formal development and social propagation are likely simultaneous aspects of linguistic change. The propagation of wanna colligated with nouns and noun phrases, without an intervening infinitive, is possibly due to casual speech in everyday exchanges; the functional shift that results in wanna be as a colloquial phrase alluding to desire and status is widely familiar. Innovation repeatedly colors speech; propagation depends on social impulses that over time and in various milieus promote few changes among many possibilities.

References


Appendix

Table 1. The occurrence throughout world Englishes of *want to* and *wanna* in three structural positions

| Frequencies by Variety | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1a. 767 176 320 63 177 96 95 16 32 42 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1b. 123 20 25 12 12 10 20 3 3 1 8 12 11 4 2 3 3 2 3 5 |
| 2a. 1127 257 968 190 400 196 176 65 71 62 120 98 98 67 108 90 71 62 42 81 |
| 2b. 25 1 5 4 5 0 4 0 0 0 1 3 1 0 0 0 0 0 0 0 2 |
| 3a. 619 166 604 117 252 114 104 39 43 36 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 3b. 6 1 1 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 |

Table 2. At the close of clauses, the vowels for *to* in *want to*, and the occurrence of *wanna*

| Frequencies by Variety | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| a. 76 19 56 16 31 23 11 5 11 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| b. 17 5 34 6 10 11 6 0 3 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| c. 2 1 0 0 1 0 1 0 0 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

Table 3. At the close of clauses, the vowels for *to* in *want to*, and the occurrence of *wanna*

| Frequencies by Variety | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| (a) 4 0 1 1 3 0 0 0 1 0 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| (b) 46 13 29 3 13 3 10 3 5 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| (c) 86 23 53 6 25 5 31 8 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| (d) 4 0 7 1 0 0 0 0 1 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

Note: The substitutability and diffusion of *want to* and *wanna* in world Englishes.
Table 4. Imperatives for *want to* and *want* in world Englishes

| Frequency by Variety | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|----------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| a.1 *want to*        | 390| 180| 290| 84 | 117| 27 | 29 | 60 | 66 | 73 | 65 | 52 | 62 | 50 | 34 | 47 | 42 | 29 |
| a.2 *want to*        | 131| 21 | 61 | 23 | 26 | 4  | 18 | 7  | 9  | 2  | 2  | 7  | 8  | 8  | 5  | 14 | 9  | 3  | 3  | 6  |
| b.1 *wanna*          | 10 | 3  | 1  | 1  | 2  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| b.2 *wanna*          | 10 | 0  | 4  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

Table 5. Patterns of *want to* and *wanna* in world Englishes as indicatives of principal clauses

<table>
<thead>
<tr>
<th>Frequency by Variety</th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. <em>want to</em></td>
<td>6300</td>
<td>1854</td>
<td>4478</td>
<td>1105</td>
<td>1734</td>
<td>899</td>
<td>1494</td>
<td>375</td>
<td>551</td>
<td>538</td>
<td>702</td>
<td>601</td>
<td>877</td>
<td>400</td>
<td>618</td>
<td>912</td>
<td>613</td>
<td>700</td>
<td>319</td>
<td>572</td>
</tr>
<tr>
<td>b. <em>wanna</em></td>
<td>1493</td>
<td>717</td>
<td>1694</td>
<td>435</td>
<td>480</td>
<td>352</td>
<td>501</td>
<td>214</td>
<td>151</td>
<td>228</td>
<td>207</td>
<td>224</td>
<td>167</td>
<td>193</td>
<td>238</td>
<td>328</td>
<td>275</td>
<td>291</td>
<td>227</td>
<td>234</td>
</tr>
<tr>
<td>c. <em>wanna</em></td>
<td>132</td>
<td>21</td>
<td>53</td>
<td>61</td>
<td>26</td>
<td>4</td>
<td>18</td>
<td>7</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>14</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>d. <em>wanna</em></td>
<td>814</td>
<td>206</td>
<td>555</td>
<td>99</td>
<td>163</td>
<td>71</td>
<td>174</td>
<td>47</td>
<td>83</td>
<td>75</td>
<td>69</td>
<td>58</td>
<td>69</td>
<td>34</td>
<td>56</td>
<td>77</td>
<td>75</td>
<td>58</td>
<td>38</td>
<td>68</td>
</tr>
<tr>
<td>e. <em>wanna</em></td>
<td>1063</td>
<td>256</td>
<td>82</td>
<td>154</td>
<td>43</td>
<td>82</td>
<td>33</td>
<td>48</td>
<td>19</td>
<td>38</td>
<td>20</td>
<td>54</td>
<td>37</td>
<td>20</td>
<td>19</td>
<td>24</td>
<td>40</td>
<td>34</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>f. <em>wanna</em></td>
<td>35</td>
<td>9</td>
<td>37</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>15</td>
<td>8</td>
<td>20</td>
<td>9</td>
<td>16</td>
<td>23</td>
<td>14</td>
<td>10</td>
<td>3</td>
<td>40</td>
<td>11</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>g. <em>wanna</em></td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>17</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>13</td>
<td>13</td>
<td>3</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>h. <em>wanna</em></td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>i. <em>wanna</em></td>
<td>1616</td>
<td>521</td>
<td>928</td>
<td>246</td>
<td>455</td>
<td>209</td>
<td>393</td>
<td>166</td>
<td>147</td>
<td>143</td>
<td>113</td>
<td>139</td>
<td>117</td>
<td>97</td>
<td>147</td>
<td>168</td>
<td>124</td>
<td>118</td>
<td>115</td>
<td>99</td>
</tr>
<tr>
<td>j. <em>wanna</em></td>
<td>256</td>
<td>69</td>
<td>152</td>
<td>33</td>
<td>59</td>
<td>42</td>
<td>50</td>
<td>14</td>
<td>19</td>
<td>27</td>
<td>24</td>
<td>25</td>
<td>16</td>
<td>14</td>
<td>21</td>
<td>34</td>
<td>22</td>
<td>18</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>a’ <em>wanna</em></td>
<td>181</td>
<td>28</td>
<td>99</td>
<td>15</td>
<td>31</td>
<td>13</td>
<td>35</td>
<td>5</td>
<td>6</td>
<td>21</td>
<td>48</td>
<td>40</td>
<td>63</td>
<td>10</td>
<td>17</td>
<td>15</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>b. <em>wanna</em></td>
<td>36</td>
<td>9</td>
<td>28</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>12</td>
<td>3</td>
<td>10</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>c. <em>wanna</em></td>
<td>13</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d. <em>wanna</em></td>
<td>6</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>e. <em>wanna</em></td>
<td>14</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>f. <em>wanna</em></td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g. <em>wanna</em></td>
<td>163</td>
<td>26</td>
<td>79</td>
<td>9</td>
<td>23</td>
<td>14</td>
<td>22</td>
<td>6</td>
<td>16</td>
<td>4</td>
<td>17</td>
<td>27</td>
<td>22</td>
<td>4</td>
<td>6</td>
<td>28</td>
<td>9</td>
<td>15</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>h. <em>wanna</em></td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
The substitutability and diffusion of *want to* and *wanna* in world Englishes

Double Modal

b. *want to*

|     | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1   | 14 | 5  | 12 | 3  | 3  | 1  | 3  | 2  | 1  | 1  | 0  | 3  | 0  | 3  | 1  | 0  | 4  | 0  | 3  | 1  |
| 2   | 41 | 12 | 36 | 11 | 19 | 13 | 6  | 1  | 4  | 6  | 5  | 5  | 4  | 4  | 3  | 0  | 2  | 3  | 4  | 1  |
| 3   | 3683| 1818| 3095 | 738| 1227| 798| 1018| 261| 283| 429| 679| 593| 597| 533| 472| 446| 328| 398| 346| 427|
| 4   | 4876| 1532| 3791 | 792| 1349| 837| 804| 222| 256| 301| 637| 501| 527| 416| 343| 293| 228| 243| 215| 264|
| 5   | 2  | 0  | 8  | 0  | 1  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 2  | 1  |
| 6   | 208 | 41  | 172 | 29 | 47 | 22 | 25 | 11 | 13 | 9  | 20 | 14 | 16 | 11 | 13 | 14 | 16 | 9  | 9  |
| 7   | 2361| 1151| 3189 | 722| 1047| 658| 833| 261| 257| 344| 422| 390| 381| 333| 387| 498| 314| 326| 268| 330|
| 8   | 3270| 971 | 3412 | 564| 1152| 563| 816| 248| 345| 237| 460| 466| 447| 235| 311| 575| 504| 373| 188| 379|
| 9   | 1488| 564 | 1049 | 233| 469| 239| 284| 94 | 79 | 152| 156| 134| 173| 156| 114| 104| 83 | 105| 104| 135|
| 10  | 1028| 241 | 972 | 181| 324| 132| 81 | 28 | 30 | 24 | 85 | 72 | 106| 42 | 47 | 32 | 29 | 37 | 18 | 38 |

b'. *wanna*

|     | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 3   | 17 | 3  | 6  | 0  | 0  | 1  | 2  | 3  | 1  | 0  | 9  | 5  | 3  | 0  | 2  | 2  | 0  | 0  | 1  |
| 4   | 84 | 8  | 28 | 9  | 12 | 3  | 9  | 1  | 0  | 4  | 12 | 28 | 15 | 4  | 3  | 12 | 1  | 1  | 2  |
| 7   | 7  | 0  | 9  | 2  | 0  | 1  | 0  | 0  | 1  | 0  | 5  | 4  | 0  | 0  | 2  | 0  | 0  | 0  | 1  |
| 8   | 6  | 3  | 17 | 2  | 6  | 0  | 1  | 1  | 6  | 1  | 5  | 8  | 2  | 1  | 5  | 5  | 4  | 3  | 1  |
| 9   | 6  | 4  | 6  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 2  | 1  | 1  | 3  | 0  |
| 10  | 12 | 3  | 11 | 7  | 12 | 1  | 2  | 2  | 1  | 0  | 3  | 3  | 9  | 0  | 0  | 5  | 2  | 0  | 3  |

VP + Object Clause
c. *want to*

|     | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1   | 2602| 755 | 2291| 372| 665| 339| 430| 148| 210| 141| 218| 247| 260| 147| 179| 351| 201| 186| 92 | 182|
| 2   | 1459| 333 | 1109| 238| 304| 180| 316| 105| 134| 108 |175| 146| 158 |127| 129| 154| 163| 160| 95 | 148|
| 3   | 1541| 520 | 1450| 290| 527| 249| 332| 112| 122| 112 |175| 157| 155 | 99| 153| 148 |105| 121| 66 | 141|
| 4   | 93  | 44  | 114 | 19 | 38 | 33 | 22 | 15 | 20 | 19 | 15 | 20 | 21 | 11 | 11 | 7  | 10 | 15 | 7  |
| 5   | 16  | 10  | 16 | 8  | 11 | 7  | 1  | 3  | 1  | 1  | 7  | 0  | 9  | 4  | 3  | 9  | 1  | 1  | 1  |
| 6   | 173 | 123 | 258 | 69 | 76 | 65 | 56 | 23 | 14 | 26 | 28 | 34 | 28 | 21 | 50 | 31 | 29 | 34 | 31 |
| 7   | 130 | 51  | 122 | 34 | 51 | 28 | 24 | 6  | 14 | 8  | 26 | 21 | 16 | 7  | 18 | 27 | 12 | 13 | 4  |
| 8   | 127 | 52  | 179 | 28 | 57 | 38 | 19 | 11 | 17 | 9  | 25 | 21 | 14 | 8  | 24 | 27 | 9  | 17 | 10 | 14 |
| 9   | 154 | 70  | 172 | 35 | 57 | 29 | 33 | 10 | 23 | 19 | 32 | 20 | 15 | 22 | 17 | 34 | 12 | 23 | 12 | 11 |
### Frequency by Variety

**c. wanna**

|   | US | CA | GB | IE | AU | NZ | IN | LK | PK | BD | SG | MY | PH | HK | ZA | NG | GH | KE | TZ | JM |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 | 84 | 6  | 32 | 11 | 9  | 4  | 15 | 5  | 1  | 2  | 29 | 13 | 21 | 6  | 10 | 6  | 11 | 5  | 5  | 1  |
| 2 | 8  | 2  | 3  | 1  | 2  | 1  | 2  | 0  | 1  | 2  | 5  | 2  | 5  | 2  | 1  | 2  | 2  | 0  | 1  | 0  |
| 3 | 23 | 6  | 9  | 7  | 4  | 1  | 3  | 0  | 1  | 2  | 1  | 2  | 3  | 0  | 1  | 2  | 0  | 1  | 0  | 4  |
| 4 | 1  | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 5 | 5  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 6 | 15 | 0  | 7  | 0  | 1  | 1  | 1  | 0  | 0  | 2  | 1  | 0  | 0  | 0  | 0  | 2  | 2  | 0  | 0  | 0  |
| 7 | 3  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 2  | 0  | 0  | 0  | 2  | 0  | 1  | 0  | 0  |
| 8 | 4  | 0  | 1  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 9 | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

Copula + Predicate Complement

### Frequency by Variety

**d. want to**

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>278</td>
<td>93</td>
<td>249</td>
<td>61</td>
<td>98</td>
<td>57</td>
<td>40</td>
<td>19</td>
<td>16</td>
<td>16</td>
<td>21</td>
<td>21</td>
<td>32</td>
<td>15</td>
<td>30</td>
<td>22</td>
<td>17</td>
<td>18</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>96</td>
<td>267</td>
<td>42</td>
<td>91</td>
<td>46</td>
<td>62</td>
<td>26</td>
<td>36</td>
<td>25</td>
<td>26</td>
<td>26</td>
<td>36</td>
<td>24</td>
<td>34</td>
<td>17</td>
<td>48</td>
<td>67</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>105</td>
<td>34</td>
<td>124</td>
<td>43</td>
<td>44</td>
<td>15</td>
<td>27</td>
<td>9</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>10</td>
<td>22</td>
<td>9</td>
<td>22</td>
<td>13</td>
<td>10</td>
<td>9</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>29</td>
<td>74</td>
<td>18</td>
<td>18</td>
<td>10</td>
<td>18</td>
<td>9</td>
<td>12</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>14</td>
<td>10</td>
<td>21</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>9</td>
<td>22</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>11</td>
<td>7</td>
<td>10</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

### Frequency by Variety

**d. wanna**

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 6. Patterns of mood for want to and wanna in world Englishes: Interrogative

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>554</td>
<td>152</td>
<td>427</td>
<td>88</td>
<td>193</td>
<td>88</td>
<td>133</td>
<td>48</td>
<td>60</td>
<td>49</td>
<td>63</td>
<td>56</td>
<td>86</td>
<td>40</td>
<td>43</td>
<td>64</td>
<td>47</td>
<td>52</td>
<td>40</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>10</td>
<td>15</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>286</td>
<td>89</td>
<td>153</td>
<td>31</td>
<td>79</td>
<td>53</td>
<td>63</td>
<td>14</td>
<td>17</td>
<td>15</td>
<td>25</td>
<td>13</td>
<td>17</td>
<td>28</td>
<td>14</td>
<td>10</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>86</td>
<td>173</td>
<td>39</td>
<td>70</td>
<td>27</td>
<td>57</td>
<td>3</td>
<td>12</td>
<td>13</td>
<td>21</td>
<td>36</td>
<td>37</td>
<td>12</td>
<td>21</td>
<td>32</td>
<td>20</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>a.</td>
<td>want to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>CA</td>
<td>GB</td>
<td>IE</td>
<td>AU</td>
<td>NZ</td>
<td>IN</td>
<td>LK</td>
<td>PK</td>
<td>BD</td>
<td>SG</td>
<td>MY</td>
<td>PH</td>
<td>HK</td>
<td>ZA</td>
<td>NG</td>
<td>GH</td>
<td>KE</td>
<td>TZ</td>
<td>JM</td>
</tr>
<tr>
<td>y/n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>123</td>
<td>20</td>
<td>30</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>21</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>47</td>
<td>7</td>
<td>16</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>a'</td>
<td>wanna</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>CA</td>
<td>GB</td>
<td>IE</td>
<td>AU</td>
<td>NZ</td>
<td>IN</td>
<td>LK</td>
<td>PK</td>
<td>BD</td>
<td>SG</td>
<td>MY</td>
<td>PH</td>
<td>HK</td>
<td>ZA</td>
<td>NG</td>
<td>GH</td>
<td>KE</td>
<td>TZ</td>
<td>JM</td>
</tr>
<tr>
<td>wh/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>5</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>24</td>
<td>70</td>
<td>11</td>
<td>32</td>
<td>19</td>
<td>20</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>13</td>
<td>12</td>
<td>16</td>
<td>6</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>129</td>
<td>56</td>
<td>136</td>
<td>31</td>
<td>61</td>
<td>23</td>
<td>33</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>17</td>
<td>21</td>
<td>21</td>
<td>15</td>
<td>18</td>
<td>20</td>
<td>21</td>
<td>13</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>402</td>
<td>139</td>
<td>346</td>
<td>96</td>
<td>138</td>
<td>88</td>
<td>146</td>
<td>64</td>
<td>72</td>
<td>44</td>
<td>48</td>
<td>66</td>
<td>41</td>
<td>47</td>
<td>54</td>
<td>82</td>
<td>51</td>
<td>52</td>
<td>33</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>8</td>
<td>15</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>55</td>
<td>11</td>
<td>36</td>
<td>8</td>
<td>12</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>5</td>
<td>13</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
### Table 7. Relative clauses for *want to* and *wanna* in world Englishes

**b. wanna**

|     | US  | CA  | GB  | IE  | AU  | NZ  | IN  | LK  | PK  | BD  | SG  | MY  | PH  | HK  | ZA  | NG  | GH  | KE  | TZ  | JM  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 4   | 10  | 1   | 13  | 1   | 1   | 0   | 3   | 1   | 1   | 0   | 3   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 5   | 1   | 0   | 0   | 1   | 2   | 0   | 1   | 1   | 1   | 1   | 2   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| 11  | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 12  | 0   | 0   | 0   | 0   | 1   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 16  | 2   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |

**Table 8. Want to and wanna in the subordinate clauses of world Englishes**

**a. subordinator want to**

|     | US  | CA  | GB  | IE  | AU  | NZ  | IN  | LK  | PK  | BD  | SG  | MY  | PH  | HK  | ZA  | NG  | GH  | KE  | TZ  | JM  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1   | 56  | 42  | 103 | 18  | 21  | 17  | 16  | 5   | 4   | 13  | 9   | 7   | 4   | 7   | 8   | 4   | 2   | 4   | 2   |
| 2   | 59  | 26  | 70  | 14  | 20  | 18  | 9   | 6   | 5   | 7   | 11  | 6   | 9   | 5   | 4   | 9   | 9   | 3   | 3   |
| 3   | 2637| 687 | 1968| 388 | 777 | 368 | 460 | 149 | 231 | 180 | 342 | 301 | 393 | 198 | 226 | 460 | 311 | 211 | 184 |
| 4   | 76  | 12  | 28  | 7   | 7   | 4   | 17  | 7   | 1   | 2   | 8   | 19  | 22  | 4   | 2   | 1   | 4   | 11  | 1   |
| 5   | 1003| 335 | 872 | 163 | 365 | 162 | 182 | 46  | 85  | 62  | 141 | 109 | 103 | 61  | 108 | 144 | 90  | 98  | 55  |
| 6   | 24471| 7393| 23984| 5599| 9241| 4978| 6538| 2275| 2896| 2958| 3098| 3132| 3131| 2543| 2506| 2673| 1846| 1802| 1739|

### Table 7. Relative clauses for *want to* and *wanna* in world Englishes
The substitutability and diffusion of *want to* and *wanna* in world Englishes

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35</td>
<td>36</td>
<td>579</td>
<td>106</td>
<td>3</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>28</td>
<td>17</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Table 9.* *Want to* and *wanna* + *be* as modifiers and noun phrase

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>5</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>1</td>
<td>20</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>549</td>
<td>92</td>
<td>274</td>
<td>54</td>
<td>125</td>
<td>35</td>
<td>86</td>
<td>34</td>
<td>45</td>
<td>23</td>
<td>87</td>
<td>107</td>
<td>100</td>
<td>28</td>
<td>22</td>
<td>69</td>
<td>31</td>
<td>26</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>32</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>34</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td>15</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>101</td>
<td>3</td>
<td>18</td>
<td>5</td>
<td>13</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

*Table 9.* *Want to* and *wanna* + *be* as modifiers and noun phrase

<table>
<thead>
<tr>
<th>modifier</th>
<th>want to be</th>
</tr>
</thead>
</table>

a. *want to be*

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
a. **wanna be**

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Noun

b. **want to be**

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

b'. **wanna be**

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>LK</th>
<th>PK</th>
<th>BD</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NG</th>
<th>GH</th>
<th>KE</th>
<th>TZ</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Chapter 4

Dialect contact influences on the use of GET and the GET-passive

Elisabeth Bruckmaier
Ludwig-Maximilians-University Munich

GET is a highly frequent and multifunctional English verb but has so far gone unnoticed in variationist studies of World Englishes. This study aims at exploring to what extent dialect contact contributes to the variation of GET in British, Jamaican, and Singaporean English, in particular to variation in the frequencies of its word-forms and in the use of the GET-passive. For that purpose, all tokens of GET in the ICE (International Corpus of English) corpora of Great Britain, Jamaica, and Singapore were analysed for form and meaning. The results demonstrate that influence from the major standard varieties British and American English as well as substrate influence can be made responsible for the variation of GET.

Keywords: GET-passive, word-forms, World Englishes, substrate influence, dialect contact

1. Introduction

GET often goes unnoticed among speakers of English, but not much less so in linguistic studies, and this contrasts decisively with its exceptional frequency and versatility. This analysis sets out to shed light on its distribution and on the dialect contact influences on its use in the standard language systems of British English, Jamaican English, and Singaporean English, with a special focus on the GET-passive. The data used mainly come from three corpora of the International Corpus of English (ICE) project, viz. ICE-Great Britain (ICE-GB), ICE-Jamaica (ICE-JA), and ICE-Singapore (ICE-SIN). In addition, data from other ICE corpora and

1. The ICE corpora comprise about 1 million words each and follow a common design of 500 texts of 2,000 words. 60% of the texts are spoken, 40% are written. The texts all date from after 1989.
from the *Corpus of Global Web-based English* (GloWbE)\(^2\) were used. ICE-GB has been chosen as a basis of comparison, with English used as a Native Language (ENL), while ICE-JA and ICE-SIN have been chosen as examples of Postcolonial Englishes (PCEs), with English used as a Second Language (ESL).

Jamaica’s sociolinguistic situation differs from Singapore’s in that the former is characterised by the presence of a post-creole continuum and the latter by a multilingual ecology. In Jamaica, a mesolectal form\(^3\) of Jamaican Creole is most widely audible in public, while full command of Standard English indicates an elitist status. In the last few decades, with a growing sense of nationalism, Jamaican Creole has also made inroads into more formal contexts (cf. Schneider 2007: 234–236). Since independence in the 1960s, the exonormative British English teaching target has eroded and American English influence has been on the increase (cf. Mair & Sand n.d.). Singapore’s language policy promotes an English-based bilingualism with English as the first language, and one of the other three official languages – Malay, Mandarin, and Tamil – as the so-called mother-tongue second language (cf. Deterding 2007: 85–86). Colloquial Singapore English, the variety spoken on an everyday basis by the majority of Singaporeans, is a contact language heavily influenced by the substrates and can be regarded as the low variety in a diglossic situation, with Standard Singaporean English the high variety used for formal purposes, such as education, law, and media (cf. Lim 2012: 285). In the present study, the focus is on the standard varieties.

Since variations in the use of *get* plainly occur, the differences observed are likely due to several influences: style and mode, the impact of second language acquisition (SLA), and, most importantly, dialect contact, which means that variation occurs because of the influence from a special use or distribution in another variety in close contact. As for the influence of British and American English on the two PCEs under study, in both Jamaica and Singapore, British English provided an initial model. Yet today, American English exercises an influence as great as British English on Jamaican English, a result of inroads by US business and media, as well as the massive increase in immigration. Indeed, British English is more a proclaimed than actual norm (cf. Mair & Sand n.d.). Lim (2012: 288) explicitly refers to increasing American English influence on Singapore, particularly

\(^2\) GloWbE contains 1.9 billion words from web pages in 20 different English-speaking countries and was released in 2013. Since the corpus exclusively contains web-based English, it cannot be considered representative of English as a whole. However, its size is a clear advantage. The British part consists of 387,615,074 words, the Jamaican one of 39,663,666, and the Singaporean subcorpus consists of 42,974,705 words.

\(^3\) “Mesolect” can be defined as any variety intermediate between the basilect and the acrolect in a post-creole continuum (cf. DeCamp 1971), such as is found in Jamaica.
Dialect contact influences on the use of GET and the GET-passive

In the domain of the media, because the majority of Singapore’s TV programmes are American productions (also cf. Lim & Ansaldo 2013b). As for substrate effects, their extent will be detailed in the respective sections.

In the following, a description of the token frequencies of get and its individual word-forms will introduce the analysis. I suggest that paying attention to the word-form can point to interesting avenues for more detailed research because particular patterns or uses can be linked to particular word-forms. In a second part, results from an analysis of the get-passive will be presented in greater detail. Dialect contact influences, i.e. not only influence from American English and different degrees of orientation towards the input variety British English, but also substrate effects are expected to contribute to the variation of get in the PCEs. More specifically, for Jamaican English a higher frequency and more unrestricted use of the get-passive than in British English can be assumed, while frequency and use are expected to be similar in Singaporean and British English. Furthermore, mode, i.e. the distinction between spoken and written language, is assumed to determine the variation of the get-passive, in that it is more frequent in spoken than in written language, in contrast to the be-passive. The conclusion will summarise the findings of the analyses and assess their significance.

2. Token frequencies of get and its word-forms

Looking at get over time reveals that probably few other verbs have attracted as much prescriptive resistance. Nevertheless, get has seen a strong increase in the history of English: research based on A Representative Corpus of Historical English Registers (ARCHER) has shown an increase in get-constructions from around 640 tokens per million words (pmw) in the mid-18th century to 2,175 tokens pmw in the second half of the 20th century. Moreover, a generally more widespread use of get in American than in British English can be assumed from the mid-19th century onwards, with the regional differences in the second half of the 20th century amounting to around 360 tokens pmw (cf. Hundt 2001: 59–60), even though prescriptivism within inner-circle varieties has been stronger in American than in British English.

The following section will clarify how World Englishes can be ranked as far as overall token frequencies of get are concerned. Possible effects of SLA, e.g. the “lexical teddy bear” effect (cf. Hasselgren 1994: 237), which is an overuse of already frequent and well-established items, lead one to expect a highly frequent use in Jamaican and Singaporean English. In a second step, the distribution of the individual word-forms of get will be explored in order to substantiate the claim that taking the word-form as the basis for analysis allows statements about
what factors are relevant for variation because individual word-forms are linked to specific uses.

2.1 Overall token frequencies of get across varieties

Figure 1 illustrates the results of a query for get in ICE-GB, ICE-JA, and ICE-SIN, as well as in the British, Jamaican, and Singaporean subcorpus of GloWbE. ICE-GB is far ahead, with 3,365 occurrences of get, while ICE-JA has only 2,502 and ICE-SIN 2,684. The differences between British English and each of the PCEs are highly significant. In GloWbE, British English again turns out to have the highest token figure, closely followed by Singaporean English, with Jamaican English last. Although the differences between British English and the other varieties are less extreme in GloWbE than in ICE, it is apparent that a more frequent use of the high-frequency verb get in ESL than in ENL, in line with an assumed “lexical teddy bear” effect, cannot be determined at this point.

![Figure 1. Token frequencies of get in ICE and GloWbE (tokens pmw)](image)

4. I use small capital letters to denote a lemma which subsumes all the inflected forms, and italics to denote a particular word-form. Thus, get subsumes get, gets, getting, got, gotta, and gotten, and have got to represents had got to as well as has got to, have got to, and having got to.

5. The numbers for ICE comprise valid tokens only, i.e. tokens that occur in analysable constructions of their own, while the results from GloWbE have not been cleaned.

6. ICE-GB and ICE-JA: \( p < 0.001, \chi^2 = 133.39, \text{ df } = 1 \); ICE-GB and ICE-SIN: \( p < 0.001, \chi^2 = 79.82, \text{ df } = 1 \).
The lower numbers in the PCEs as compared to British English are striking and are confirmed by a simple frequency search in several ICE corpora. As Figure 2 illustrates, in the case of get no ESL corpus exceeds the token frequencies of ICE-GB. It might be that prescriptivism against the use of get is more effective in outer-circle varieties than in British English. In any case, the results presented so far allow the conclusion that the higher overall token figure of get in ICE-GB, as compared to ICE-JA and ICE-SIN, reflects a real regional difference in the absolute frequency of use. British English is ahead in the use of get, while outer-circle varieties feature lower absolute token frequencies. However, there might still be individual constructions that are overused in the PCEs, as compared to British English. The analysis of get-passives below will shed further light on this.

![Figure 2. Token frequencies of get in ICE (absolute token numbers)](image)

### 2.2 The distribution of the word-forms of get across varieties

As yet, it is open to question whether, in the case of get, particular patterns or uses can be linked to particular word-forms. Most researchers so far have glossed over differences between inflectional forms and have either lumped together different word-forms under the heading of lemma or have left the actual word-form that was the object of study unspecified, or have used the infinitive form (cf. Gries 2011: 238). In fact, uneven distributions of word-forms of get across ICE, particularly of got and gotten, indicate strong variability in World Englishes.

The following corpus-driven analysis is based on the assumption that the word-form has to be the basic unit of analysis and not the lemma, let alone larger

---

7. Tokens have not been cleaned for this analysis.
syntactic structures, because each word-form has its own grammar and meanings (cf. Biber 2009: 278). While a radical corpus-driven approach has more often been proclaimed than actually applied, the emphasis it places on the individual word-form has triggered new analyses. Sinclair (1991: 105) points to the importance of the individual word-form, which should become the focal point in the presentation of data. Newman and Rice suggest the notion of an “inflectional island” because

[...] syntactic/semantic properties tend to inhere in individual inflections of a verb in a register-specific manner. Furthermore, these properties may not extend across all the inflections to characterize the lemma as a whole. For us, the notion of a dictionary entry based on a lemma is still inadequate.

(Newman & Rice 2006: 255)

One should aim at choosing that degree of granularity that provides the most insightful results (cf. Gries 2011: 237). In a given case, this may be the level of inflectional forms. In the end, in a bottom-up fashion, appropriate data must be adduced because “corpora provide data on many different levels of hierarchical organization, and not all hierarchical levels are necessarily suited equally well to all tasks” (Gries 2011: 238). The following analysis of the distribution of individual word-forms of *get* across varieties of English shows major differences among them. Table 1 provides the results of a query for all word-forms of *get* in ICE-GB, ICE-JA, and ICE-SIN. It reveals that across ICE, a gulf opens up between ICE-JA and the other two corpora.8 Post-hoc tests show a particularly frequent use of *get*, as well as a particularly infrequent use of *got*, in ICE-JA compared to ICE-GB and ICE-SIN. The differences between ICE-GB and ICE-SIN do not attain statistical significance, a result which is illustrated by the very similar distribution in British and Singaporean English in Figure 3, where Jamaican English stands out.

<table>
<thead>
<tr>
<th>Table 1. Token frequencies of all word-forms of <em>get</em> in ICE-GB, ICE-JA, and ICE-SIN (absolute token numbers)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICE-GB</strong></td>
</tr>
<tr>
<td><em>get</em></td>
</tr>
<tr>
<td><em>gets</em></td>
</tr>
<tr>
<td><em>getting</em></td>
</tr>
<tr>
<td><em>got</em></td>
</tr>
<tr>
<td><em>gotta</em></td>
</tr>
<tr>
<td><em>gotten</em></td>
</tr>
<tr>
<td><strong>total get-tokens</strong></td>
</tr>
</tbody>
</table>

8. ICE-JA and ICE-GB: $p < 0.001$, $\chi^2 = 381.36$, df = 3; ICE-JA and ICE-SIN: $p < 0.001$, $\chi^2 = 271.35$, df = 3.
Dialect contact influences on the use of GET and the GET-passive

Figure 3. Distribution of the word-forms of get in ICE-GB, ICE-JA, and ICE-SIN (percentages)

Comparisons of speech-writing ratios per corpus and word-form show that got is the word-form with the highest ratio in ICE-GB (7.5 : 1) and ICE-SIN (3.9 : 1), while gets has the highest ratio in ICE-JA (4.1 : 1). That is, got is characteristic of spoken British and Singaporean English and strongly responsible for the lower absolute token number of all word-forms of get in Jamaican English as compared to British English: ICE-JA spoken has more than 1,000 got-tokens fewer than ICE-GB spoken. The lower frequency of got in Jamaican English is likely due to the infrequent use of possessive have got and of semi-modal have got to. The significant differences between Jamaican English and the other two varieties suggest that neither the use of possessive got nor the use of semi-modal gotta, both of which would indicate strong American English influence, can compensate for the non-use of have got and have got to. The difference between ICE-GB and ICE-SIN is not so easily attributable to a single word-form: while in written language only get is less frequently used in ICE-SIN than in ICE-GB, in spoken language, all word-forms except the low-frequency forms gotta and gotten yield fewer token numbers in ICE-SIN than in ICE-GB.

In the analyses above, gotta has been treated as a word-form in its own right. It is of course a coalescence of got and cliticised to, a reduction with an American English flapped pronunciation of the phoneme /t/. It is striking that while the search for gotta yields practically no results for British English, a total of 17 valid tokens occur in ICE-JA, and 11 in ICE-SIN (cf. Table 1). These results suggest a spread of gotta in Jamaican English, a finding likely due to strong American influence. For Singaporean English as well, a certain, though weaker, tendency to use
a flapped pronunciation is suggested by the results. This ties in with an increasing American English influence on Singaporean English.9

An alternation between the two past participle forms *got* and *gotten* occurs in perfect tenses in American English, but not in British English. According to Burchfield, “[n]othing points more clearly to the North Americanness of a person than the ability to use the p.a.pple [past participle] forms *got* and *gotten* in a natural manner” (1996: 338). The presence of *gotten* in a text rules out British origin (cf. Mair 2007: 457), but there is a fair amount of variation in American English, where it has been retained. American English conversation is “unexpectedly conservative in retaining the irregular form *have gotten*”, according to Biber et al. (1999: 399).

In general, *gotten* can be used with the dynamic meaning of ‘received’ or ‘obtained’, i.e. in a monotransitive or ditransitive construction, or in copular constructions with an obligatory adverbial expressing movement, or with an adjectival complement. Only *got* can be used even in American English when the meaning is stative, i.e. ‘possessed’. Burchfield points out that “[t]he matter cannot be neatly resolved since *get* has so many other senses and applications, and the prospect of examining the distribution of *gotten* and *got* in all of them would be very daunting indeed” (1996: 338). The electronic *World Atlas of Varieties of English* (eWAVE) (cf. Kortmann & Lunkenheimer 2013) provides information about two uses featuring the word-form *gotten*. Neither the typically American English use of *gotten* for dynamic meanings nor the use of *gotten* for stative meanings occurs in Jamaican English (cf. Sand 2013), Jamaican Creole (cf. Patrick 2013), or Colloquial Singapore English (cf. Lim & Ansaldo 2013a). However, the *Dictionary of Caribbean English Usage* reports that

> [t]he widely current use of *got* in these [possessive and modal] functions prompts many educated CE [Caribbean English] speakers to use AmE [American English] *gotten* in place of *got* as the pa.t. [participle] of *get*, and this is now much seen in writing. (Allsopp 2003: 264)

Contrary to the statements in eWAVE but in line with Allsopp (2003), 36 tokens of *gotten* occur in ICE-JA. In ICE-SIN, *gotten* still occurs 12 times, while it is unsurprising to find only 2 tokens of *gotten* in ICE-GB. Of all 36 tokens of *gotten* in ICE-JA, exactly half are used in monotransitive constructions expressing an event of receiving or obtaining, 7 occur in particle verbs (PVs), 6 in copular constructions with an adjectival complement, 4 in motion constructions with an obligatory adverbial, and 1 in a complex-transitive construction. Of all 12 tokens in ICE-SIN, just as in ICE-JA, exactly half are used in monotransitive constructions, 5 occur in
PVs, and 1 in a motion construction. About three quarters of all instances of gotten occur in spoken language in both varieties. This means that gotten is virtually not used in British English, while it occurs quite frequently in Jamaican English. It occurs to a limited extent in Singaporean English, where one can assume that colloquialisation and orientation towards the British English model work in opposite directions, ultimately leading to a moderately high frequency of gotten. The most frequent uses of the form occur in monotransitive constructions and PVs.

To sum up, it has been shown that taking the word-forms of get as the basis for analysis allows first claims about reasons for the variation it displays. This is because individual word-forms can be linked to particular uses.

3. The get-passive

In the following, one of the structures into which get enters, viz. the get-passive, will be analysed in greater detail in order to trace further and more specific dialect contact influences. In the first section, the nature of the passive in general will be outlined, with a focus on the get-passive. In a second step, factors that influence variation in the get-passive will be explained in detail and corresponding hypotheses will be formulated. The special meaning and use of the get-passive, attributable to its historical development involving a grammaticalisation process, will be described in a separate section before a summary of the hypotheses is provided. Finally, the results from the data will be presented.

3.1 The nature of the (get-) passive

The be- and the get-passive, as contrasted with the active voice, have in common the rearrangement of two clause elements – viz. of the active subject, which becomes the passive agent, and of the active object, which becomes the passive subject – and the addition of by before the agent. The relations of meaning are usually the same in the active and passive voice, while the emphasis put on certain elements changes. For instance, the agent receives end-weight in the passive. Most often, however, the agent is omitted altogether. In some cases, the change in the order of elements may also lead to a change in the propositional content of the sentence (cf. Quirk et al. 1985: 159–165, 1390).

The usual formal definition of the get-passive is get + past participle (cf. e.g. Quirk et al. 1985: 167). Past participles can oscillate between an adjectival and a verbal use, but in order to be able to classify and quantify data, certain cut-off points need to be determined and adjectival past participles need to be distinguished from verbal past participles. For purposes of classification, I follow Quirk
et al. (1985: 160–171) in their tripartite division of combinations of get + past participle into central passives, semi-passives, and pseudo-passives:

- central passives (participle is verbal) = “get-passives”, e.g. *He got arrested.*
- semi-passives (participle is both verbal and adjectival), e.g. *He got interested in linguistics.*
- pseudo-passives (participle is adjectival) = copular use, e.g. *He got bored.*

Boundaries between the three categories are fuzzy, and many examples of get + past participle are ambiguous: in actual data, clues concerning the interpretation will often only be implicit and any analysis of get-passives is very much dependent on the close interpretation of the whole context. Only combinations of get and unambiguous verbal participles will be termed “get-passives” in the following analyses. An example from the data is *You know your bikes will get stolen* (<ICE-SIN:S1A-009#235:1:A>): it could easily be expanded with a *by*-agent, it has a correspondence with an active clause, and the participle cannot be premodified by *very*. Note that the presence of a *by*-phrase, which often makes a get + past participle construction verbal (for details cf. Quirk et al. 1985: 168–169), is particularly rare for get-passives because they are preferably used to emphasise the subject referent’s unfavourable condition.

Previous analyses have treated the get-passive only marginally and have ascribed to it the role of a “modern variant of the be-passive” (Hübler 1998: 162). However, it displays typical characteristics, which can be summarised as follows:

- low frequency
- usually no expressed animate agent
- only with dynamic verbs
- implication of responsibility on the part of the subject
- expression of adversity or benefit
- informal register

First, the get-passive is much less frequent than the be-passive (cf. Quirk et al. 1985: 161). Biber et al. state that the get-passive is “extremely rare” (1999: 476), making up only 0.1% of all verbs even in conversation. Second, there is usually no expressed animate agent in get-passives. Thus, the examples *James got beaten last night* and *The cat got run over by a bus* are more likely to occur than *James got caught by the police* (cf. Quirk et al. 1985: 121–127, 161). Third, Huddleston and Pullum (2002: 1442) point out that get-passives are found only with dynamic verbs, which can be explained by the fact that other uses of get are always dynamic (except, it should be added, in cases such as *I got a problem*, where *got* = ‘have’). Fourth and fifth, responsibility on the part of the subject and the expression of adversity or benefit are two further characteristics pointed out by Huddleston and
Pullum (2002: 1442–1443), as in *She managed to get transferred to the finance department* or *My watch got stolen*. Emphasis is on the preferably human subject, who is adversely affected by the event in many cases. When an inanimate subject does occur, “*some* human associated with the subject, or with the event in some other capacity, may either retain responsibility, be emotionally involved, or adversely affected” (Givón 1993: 69), as in *How did this window get opened?*, where the speaker wants to convey an unfavourable attitude towards the action and the hearer is held responsible (cf. Quirk et al. 1985: 161; Givón 1993: 68–70). Finally, the get-passive is said to be avoided in formal style (cf. Quirk et al. 1985: 161). In contrast to the be-passive, it is used more frequently in spoken than in written texts (cf. Biber et al. 1999: 476).

Empirical studies have attested to an increase in get-passives in both British and American English, with American English in the lead. Hundt (2001: 71–72), for instance, detects a substantial increase in the number of get + past participle in British English for the hundred years between 1850 and 1950, while in American English the increase started as early as 1750. For the time period between the 1960s and 1990s, Leech et al. (2009: 156) have claimed an increase in get-passives for written American English and, to a slightly lesser extent, for written British English. In the following, I will point out which factors can influence variation in the use of the get-passive. Corresponding hypotheses will then be tested with the help of the data.

### 3.2 Factors influencing the variation of the get-passive

#### 3.2.1 Prescriptivism

The get-passive is generally an infrequent phenomenon compared to the be-passive and one major reason relates to prescriptive objections against its use, particularly in more formal, written registers. These prescriptive objections do not seem to apply to all varieties to the same degree, however. In fact, it has been claimed that resistance to the use of the get-passive has generally been greater in British than in American English (cf. Denison 1998: 182) and Hundt (2001: 71) attributes the slower and later spread of get + past participle in British English to exactly this factor. While prescriptivism might be waning in general in all varieties, resulting in more get-passives, researchers suggest that today American English is leading the change towards more get-passives (cf. Leech et al. 2009: 156). Given the influence of American English on Jamaica, one can expect prescriptivism to play less of a role for get-passives in Jamaican English, which makes me expect a larger number of get-passives in this variety than in British English.
3.2.2 Colloquialisation

Colloquialisation is a phenomenon that has affected the English language in recent years, making written registers more speech-like (cf. Collins & Yao 2013). Singaporean English seems to be advanced in colloquialisation, at least among outer-circle varieties. The conservative character of British English, as far as the get-passive is concerned, has been attested by Leech et al. (2009: 156), Biewer (2009: 372–373), and Collins & Yao (2013: 495–496). If one can assume that Singaporean English is indeed more colloquialised, ICE-SIN can be expected to exceed ICE-GB in the frequency of get-passives.

3.2.3 Substrate influence

Moreover, the substrate languages must be considered in terms of their influence on standard language use in the PCEs. It is therefore necessary to take into account how Jamaican Creole and Colloquial Singapore English express passive meaning. Jamaican Creole (cf. Farquharson 2013b) has the following four possibilities, which will be explained below:

- be-passive
- get-passive
- passive without verbal coding (also called “basic passive”)
- active construction with impersonal subject

get is a common passive auxiliary in many English-based pidgins and creoles (cf. Haspelmath 2013: 358), also in the Caribbean region (cf. Allsopp 2003: 253), which leads one to expect high numbers of get-passives in Jamaican English. Also common in creoles is a passive without verbal coding, the so-called “basic passive”. Thus, apart from the be-passive, in Jamaican Creole, there is a get-passive construction, as in Di fuud get kuk ‘The food was cooked’, and a passive without verbal coding, as in Di chrii kot ‘The tree was cut’ (cf. Farquharson 2013a: 86; Farquharson 2013b). A common alternative to these constructions is the use of an active construction with an impersonal subject, as in Dem kil di kou ‘The cow was killed’ (cf. Farquharson 2013a: 86).

Winford (1993: 140–142) points to the different conversational implicatures of the be- and the get-passive in Caribbean English Creoles and also to the possibility of removing voice ambiguity by inserting get in basic passives that have an animate subject, as in Plentii soldja kil laas nait vs. Plentii soldja get kil laas nait. He remarks that

[…] CEC [Caribbean English Creole] speakers seem to employ get passives in cases where vagueness or misinterpretation might result from the use of basic passives with animate subjects. The two constructions therefore appear to be
complementary rather than competing alternatives. This suggests, incidentally, that there is no need to think of get passives as recent developments or replacements for the basic passive, as Alleyne (1987: 74) seems to think. Rather, both structures appear necessary to complement each other, and may well have developed simultaneously. (Winford 1993: 143)

I expect substrate influence to exert indirect influence on the use of get-passives in Jamaican English. That is, while for instance no passives without verbal coding are expected in standard language use, it is probable that substrate patterns lead to a more frequent and less restricted use of get-passives in Jamaican English than in British English. This occurs for three reasons: first, in Jamaican Creole, the be-passive is rarer than the get-passive; second, the get-passive is firmly established as a separate structure; and third, get is inserted in the alternative passive structure of Jamaican Creole, viz. the passive without verbal coding, to remove ambiguities.

In Colloquial Singapore English, the passive seems to be more often morphologically marked than in Jamaican Creole. The following passive constructions are used in Colloquial Singapore English (cf. Lim & Ansaldo 2013b):

- be-passive
- get-passive
- kena-passive
- give-passive

The be-passive can be used as in Standard British English, but Bao and Wee (1999: 1) note that omission of be is also possible, which would correspond to a passive without verbal coding. Like the be-passive, the get-passive is acquired through successful SLA, and no further special meaning or restriction on the get-passive is noted in the literature, with local languages exhibiting no corresponding pattern. Lee et al. (2009: 310) assume that the original properties of the get-passive are maintained in Colloquial Singapore English and demonstrate that, as in British English, the verb is eventive and affects the subject in some way.


The give-passive, as in John give his boss scold 'John was scolded by his boss', is a calque on Hokkien and is special in that the subject of the active clause, viz. The boss in The boss scolded John, does not become an oblique-marked phrase but the object of give and is obligatorily present (cf. Haspelmath 2013: 359). Like the kena-passive, the give-passive requires an adversative reading. The subject must be
animate and the lexical verb must always be in the base form (cf. Bao & Wee 1999). According to Lim and Ansaldo (2013a), the *give*-passive is a pervasive feature in Colloquial Singapore English. However, the *kena*-passive is still more common than the *give*-passive, regardless of the ethnic origin of the speaker (cf. Bao & Wee 1999). Bao and Wee (1999) claim that this is attributable to the continuing pressure of Standard English, which promotes the more standard-like form and weakens the effect of substrate influence.

I follow previous research in assuming that the *kena*- and the *give*-passive, which are used in Colloquial Singapore English to express adversity, are too marked to be used in Standard Singaporean English. Thus, I claim that in the case of passives, ICE-SIN will be oriented towards the exonormative model and will display a frequency and use of *get*-passives similar to that in British English to express an adverse effect on the patient.

3.2.4 **Mode**

It is well known that the use of *be*-passives is dependent on register and style (cf. Quirk et al. 1985: 166). While in the major reference grammars little information is provided concerning the distribution of *get*-passives as contrasted with *be*-passives specifically, I assume major differences between the two passives, with the *get*-passive being more frequently used in spoken than in written language because of its informal and colloquial character. As such it stands in clear contrast to the *be*-passive, for which the reverse holds true (cf. Biber et al. 1999: 476). Since research has suggested an advanced stage of Singaporean English in terms of colloquialisation, the gap between spoken and written language in the use of the *get*-passive can be assumed to be quite small in this variety.

3.3 The special meaning and use of the *get*-passive

In order to understand the present-day meaning and use of the *get*-passive, it is necessary to adopt a historical perspective. The *get*-passive is the result of a process of grammaticalisation:10 *get* has turned from a full verb with a possessive meaning element into an auxiliary by undergoing processes of metaphorisation and semantic bleaching. Givón (1993: 65–66, here simplified; also cf. Givón & Yang 1994) sketches a possible grammaticalisation chain for the passive meaning of *get*-constructions:

---

10. This means that a full lexical item turns into a grammatical element.
- monotransitive lexical ‘obtain’ > bitransitive locative ‘move’ (He got a horse > He got a horse for himself > He got a horse for her > He got the horse to her > He got the horse to the barn) (14th-16th century)
- ‘move’ > causative (He got her into the house > He got her to go into the house > He got her to go > He got her to play)
- causative > de-transitivisation (He got her to admit it > He got her to be admitted > He got himself to be admitted > He got to be admitted > He got admitted) (18th century)

Givón and Yang add that “[w]hile we have identified a plausible linear progression in the rise of the get-passive, we feel that it would be misleading to describe the process in purely linear terms” (1994: 145). Hundt (2001: 55) also points out that the process of grammaticalisation has to be conceived of as a multi-stranded development. If one wanted to extend the chain, however, get-passives with inanimate subjects and/or with by-phrases can be seen as further steps in the process of grammaticalisation. Necessarily, an inanimate subject takes on the role of an affected patient and cannot carry responsibility for the event described in the same way an animate subject can: consider Dixon’s (2005: 359) example That meeting got postponed, where the adversity of the situation is obvious but where the inanimate subject clearly cannot carry responsibility. Similarly, the presence of a by-agent means that the subject has less responsibility and that the middle meaning of the get-passive becomes weaker. A more grammaticalised use of the get-passive thus means a less restricted use, or, put differently, an extension of meaning.

Contextual clues, such as the use of certain verbs (e.g. want to, try, avoid) in the vicinity of get or the use of if (cf. Hundt 2001), by contrast, indicate subject responsibility and therefore retention of middle meaning. It will be of interest to see whether a more frequent use of the get-passive is accompanied by an extension of meaning or not. For Jamaican English, I expect a quite unrestricted use because of substrate influence. By contrast, I expect the use and meaning in Singaporean English to be similar to that in British English because the passives used in Colloquial Singapore English are too marked to have an influence on standard language use. In Standard Singaporean English, the get-passive is the only construction that remains for expressing adversative meaning.

3.4 Summary of the hypotheses

The hypotheses relating to the get-passive can be summarised as follows:
- There will be a higher frequency of the get-passive for Jamaican English than for British English due to influence from American English and from the substrate on Jamaican English. Similar frequencies will be found for British and
Singaporean English because the exonormative orientation of Singaporean English will limit the influence of Colloquial Singapore English and its marked passives on Singaporean English.

- The frequency of the get-passive will be higher in spoken than in written language. However, the gap between spoken and written language will be quite small in Singaporean English.
- There will be a less restricted use of the get-passive in Jamaican English in comparison with British English due to substrate influence on Jamaican English. The use and meaning of the get-passive in Singaporean English will be similar to that in British English due to the exonormative orientation of Singaporean English.

3.5 Results from the data

3.5.1 Frequencies

If one looks at absolute token numbers in the three ICE corpora (cf. Figure 4), the hypothesis concerning the higher frequency of get-passives in Jamaican English than in British English and similar frequencies in British and Singaporean English can be confirmed. ICE-GB contains 63 tokens, ICE-JA has 100, and ICE-SIN has 74. get-passives are significantly more frequent in ICE-JA than in ICE-GB ($p < 0.01$, $\chi^2 = 9.16$, df = 1) and I propose two reasons for this result. First, Jamaican English is in close geographical vicinity to the US and known to be influenced by American English, which is at the forefront of a change towards more get-passives and is characterised by a relatively low degree of prescriptivism against get. Given the close contact between the two countries, influence from American English is an obvious explanatory factor for the highly frequent use of get-passives in Jamaican English. Second, substrate influence from Jamaican Creole might reinforce the use of get-passives in Jamaican English.

![Figure 4. get-passives in ICE (absolute token numbers)]
In ICE-JA, 3.8% of all uses of get are get-passives, while in ICE-SIN, the percentage is 2.7, and in ICE-GB only 1.8. If one considers semi-passives as well, ICE-JA is set apart even more from the other two corpora, with 4.8%, as compared to 3.2% and 2.7% in ICE-SIN and ICE-GB, respectively. This means that the get-passive contributes more fully to uses of get in Jamaican English than in British English, with the difference between ICE-GB and ICE-JA being statistically highly significant \((p < 0.001, \chi^2 = 25.28, df = 1)\). The frequency difference between Singaporean and British English is not statistically significant, a result attributable to the fact that the British English exonormative model and its conservative prescriptivism against get-passives retain their influence on Singaporean English.

### 3.5.2 Mode

When one counts instances of get-passives separately for all written and spoken subcorpora, one can see a clear effect of mode: the use of the get-passive is much preferred in spoken language, despite the formal nature of passives in general. The hypothesis relating to the higher frequency of the get-passive in spoken language than in written language can therefore be confirmed without reservation. Figure 5 shows that in ICE-JA spoken, there are 12.8 get-passives per 100,000 words, in ICE-SIN spoken 9.5, and in ICE-GB spoken 8.3, while the figures for the written subcorpora are 4.6, 3.9, and 2.4, respectively.

While all mode differences within ICE are very marked and statistically highly significant,\(^\text{11}\) the difference between the spoken and the written subcorpus is most marked for British English, with a ratio of 3.5 : 1, while the PCEs exhibit lower ratios (ICE-JA: 2.8 : 1, ICE-SIN: 2.4 : 1). Note that the speech-writing ratio is higher for get-passives than for get in general in British and Jamaican English, whereas this ratio is lower for get-passives than for get in general in Singaporean English. This means that, in spoken British and Jamaican English, the get-passive is over-used in comparison to other constructions into which get can enter, while it is less frequently used in Singaporean English. Hence, the get-passive can be said to be particularly characteristic of spoken language in British and Jamaican English, but less so in Singaporean English.

In ICE-SIN, the difference between written and spoken language is smallest across all corpora and the number of get-passives in written language is comparatively high. While for ICE-GB and ICE-JA, the shares of get-passives of all tokens of get in the respective subcorpora are higher in spoken than in written language, in ICE-SIN the percentage is higher in written (3.1%) than in spoken

---

11. get-passives in ICE-GB spoken and ICE-GB written compared to the total number of words: \(p < 0.001, \chi^2 = 15.2, df = 1\); ICE-JA spoken and ICE-JA written: \(p < 0.001, \chi^2 = 17.37, df = 1\); ICE-SIN spoken and ICE-SIN written: \(p < 0.001, \chi^2 = 10.97, df = 1\).
Elisabeth Bruckmaier (2.6%) language. This means that the very informal get-passive is firmly established even in written language in Singaporean English. Since colloquialisation means that written language becomes more speech-like, a high degree of colloquialisation in Singaporean English can be confirmed. To sum up, the get-passive is a phenomenon of spoken language in all the varieties of English analysed here, and even more characteristic of spoken language than other uses of get in British and Jamaican English. Singaporean English is special in displaying an established use of the get-passive even in written language, resulting from the variety’s more advanced degree of colloquialisation as compared to British and Jamaican English. All hypotheses relating to mode can be fully confirmed.

3.5.3 Use and meaning

Of the three corpora, ICE-JA has the highest number of nonstandard examples, 16 in total, all occurring in the spoken subcorpus, e.g. Him get booed out (<ICE-JA:S1A-026#239:1:A>), [... and on Sundays you can get call out (<ICE-JA:S1B-078#X105:1:B>), with a base verb in the place of the past participle, or I don’t believe it was a one time and you get catch (<ICE-JA:S1A-061#307:1:A>), showing get not only followed by a base past participle but also with past time reference. The high number, as well as the types, of nonstandard features can be explained by influence from basilectal Jamaican Creole. The hypothesis concerning the less restricted use of the get-passive in Jamaican English can be confirmed. As has been shown above, the get-passive is very frequent in Jamaican English and constitutes a high percentage of all get-constructions in ICE-JA. The differences between this variety and both British and Singaporean English are statistically significant. It has also been shown that there are more nonstandard uses connected with the

Figure 5. Mode distribution of get-passives in ICE (tokens per 100,000 words)
The generally more unrestricted use of the get-passive in Jamaican English does not entail an extension of meaning, however. In the following, the meaning of the get-passive in the data will be presented in greater detail. For that purpose, get-passives were categorised according to whether they exhibit clear signals of a meaning close to the prototypical be-passive – i.e. feature contextual clues such as inanimate subjects or by-phrases, as in *Dr Arthur Beng said that the PAP needs the party whip otherwise it might get toppled by its own back-benchers* (<ICE-SIN:S2B-047#100:1:A>) – or whether they exhibit clear signals of middle meaning – i.e. co-occur with verbs such as *want to*, *try*, or *avoid*, as in *No I don’t want to get drawn in that at all* (<ICE-JA:S1B-050#91:2:B>), or with *if*-clauses, as in *I mean what’s the good of being dead if you’re still trying to get published on earth* (<ICE-GB:S1B-026#250:1:D>). The first category will be called “be-passive-like get-passives” in the results given below. A higher number of them indicates an extended meaning. The second category will be called “prototypical get-passives”. The numbers of get-passives without any indications were added to those of prototypical get-passives because elements of middle meaning can be considered to be standardly present
in get-passives if no other signals are used. Figure 6 provides the results in detail. It shows that be-passive-like get-passives are less frequent in ICE-JA (26%) than in the other corpora. Thus, the high frequency of get-passives in ICE-JA is not accompanied by an extension of meaning.

Furthermore, Table 2 shows that ICE-JA is the corpus where the participle standing after get most frequently expresses something negative happening to the subject. For this analysis, the participles were rated as either negative (e.g. caught, killed) or not negative/neutral (e.g. sprayed, offered) (cf. Lindquist 2009: 135–136). An extension of meaning would have to be reflected in quite the opposite distribution, viz. with Jamaican English exhibiting a much larger percentage of neutral uses where get does not express adversity. The general claim that the get-passive expresses adversity can be confirmed for all varieties. The adversity meaning seems to be particularly strong in Jamaican English, as the differences between ICE-JA and the other two corpora are statistically significant.12

| Table 2. Negative verbal participles in get-passives (percentages of all get-passives) |
|---------------------------------|-------|-------|-------|
| ICE-GB       | ICE-JA | ICE-SIN |
| negative verbal participles in |
| get-passives | 57.1%  | 73.0%  | 54.1%  |

In ICE-SIN, the distribution of meaning is very similar to that in ICE-GB: there are neither significant differences in meaning (cf. Figure 6) nor in the use of negative verbal participles (cf. Table 2), which means that the get-passive shows no signs of extended meaning in Singaporean English. I assumed that the kena- and give-passives of Colloquial Singapore English are too marked to be used in standard language, given their conspicuous structures. A search for kena in ICE-SIN shows that the kena-passive indeed comes up only twice in the whole corpus. The give-passive does not contain foreign words, but its syntax is so striking that it can be safely assumed that the number of tokens in ICE-SIN is very low if it occurs at all. For Singaporean English, a use and meaning of the get-passive similar to that in British English can be concluded: the exonormative British English model seems to be in place. As a result, the hypotheses with respect to the meaning and use of the get-passive in varieties of English can be confirmed.

12. ICE-JA and ICE-GB: $p < 0.05$, $\chi^2 = 4.39$, df = 1; ICE-JA and ICE-SIN: $p < 0.01$, $\chi^2 = 6.71$, df = 1.
4. Conclusion

The above analysis has attempted to shed light on variation in the use of get in British, Jamaican, and Singaporean English, paying particular attention to dialect contact influences. An analysis of the frequencies of get and its individual word-forms introduced the analysis. It has been shown that Jamaican and Singaporean English, as well as all other PCEs analysed, feature lower token numbers of get than British English. This is striking and runs counter to an expected overuse of versatile high-frequency verbs in ESL. However, in the frequencies of individual word-forms, a very similar distribution for British and Singaporean English was detected, confirming an overall British English model for Singaporean English. By contrast, Jamaican English is differentiated from the other varieties mainly through a much lower frequency of got. This can be associated with an avoidance of the distinctly British English constructions have got and have got to, i.e. the possessive and semi-modal uses of get, which are also avoided in American English. In both the Jamaican and the Singaporean corpus, the word-forms gotta and gotten occur, which suggests increasing American English influence on both PCEs. The uneven distributions of get and its word-forms across ICE have indicated strong variation in World Englishes and have shown that it is worth attending closely to the individual word-form.

The analysis of get-passives has revealed that Jamaican English displays the highest token frequency of get-passives and British English the lowest, with Singaporean English covering the middle ground. As reasons for the clear difference between British and Jamaican English I suggest influence from American English and from the substrate on Jamaican English, two factors that promote the number of get-passives. In Singaporean English, by contrast, a greater degree of colloquialisation coincides with the exonormative British English model, so that only slightly more get-passives occur than in British English.

Substrate influence of Jamaican Creole on the use of get-passives in the speech of more literate Jamaicans is apparent in nonstandard get-passive use, e.g. in constructions with missing or nonstandard subjects, nonstandard time reference, missing auxiliaries, and base forms in place of inflected verbs. Together with their high frequency and a more flexible use of person and tense, this points to a less restricted use of get-passives in Jamaican English, which, nevertheless, does not entail an extension of meaning towards be-passive-like get-passives. In Singaporean English, the get-passive is used in much the same way as in British English, only more frequently. Both the kena-passive and the give-passive are used in Colloquial Singapore English to express adversity, but are not used in educated speech. Apparently, in Standard Singaporean English, the get-passive performs the functions that the kena- and the give-passive perform in Colloquial Singapore
English. Colloquialisation and orientation towards the British English standard act as counteracting factors leading to a medium frequency of the get-passive and a use similar to that in British English. As far as mode is concerned, the get-passive is a phenomenon of spoken language in all varieties, but the gap between spoken and written language has been shown to be smallest in Singaporean English, in line with advanced colloquialisation.

To sum up, in the analyses of the frequencies of get and its word-forms and of the get-passive, no effects of SLA could be detected. By contrast, mode differences, influence of the major standard varieties and substrate influence have been traced to an appreciable extent, as was hypothesised. It has thus been demonstrated that dialect contact plays a significant role in shaping the use of get in the PCEs considered here. In the case of Jamaican English, American English influence as well as substrate influence are apparent in the use of gotta and gotten, the low frequency of have got (to), and the high frequency and unrestricted use of get-passives. In the case of Singaporean English, a general orientation towards the British English model combined with colloquialisation and limited but growing American English influence seem to be the major factors at work leading to a distribution of word-forms similar to that in British English and a moderately high frequency of gotten and get-passives.

References


CHAPTER 5

Future time marking in spoken Ghanaian English

The variation of will vs. be going to

Agnes Schneider
University of Freiburg

This article investigates outcomes in the process of structural nativization in the evolution of New English varieties in the domain of future time marking, analyzing the constraints on variation of WILL and BE GOING TO in spoken Ghanaian English (GhE) as compared to British English (BrE), using mixed effects logistic regression models. The analysis shows that in spoken GhE future time markers are not as clearly distributed syntactically, semantically, and pragmatically as in spoken BrE, which reflects its learner variety history and its status in a highly multilingual society. However, it is also shown that different future marker-verb collocations in spoken GhE may reflect first nativization processes in the variety, which corroborates previous findings that innovations in New Englishes start at the lexico-grammatical level (cf. Schneider 2007: 86–88).

Keywords: Ghanaian English, World Englishes, corpus linguistics, mixed effects logistic regression model, nativization, future time marking

1. Introduction

This paper analyzes variation in the future time system of spoken Ghanaian English (GhE), a West African New English variety. It investigates differences in the patterns of the variable system between spoken GhE and spoken British English (BrE) on the basis of corpora of private spoken conversations, adopting variationist, corpus-linguistic methods. The aim of this article is to investigate outcomes in the process of structural nativization in the evolution of New English varieties as suggested in Schneider’s (2003, 2007) Dynamic Model, and to tackle the way in which English is transformed into a new variety.

While the past decades have witnessed an increased interest in the study of the grammar of New Englishes, analyses into the variable systems of these varieties
have attracted attention only more recently. As Huber (2012b) notes, unlike in the areas of phonology and the lexicon, structural nativization in the area of morphology in educated varieties of New Englishes is not so much characterized by radical departures from the norm but by more subtle innovations which can only be grasped statistically. He states that,

[1] In transforming English into a new variety in the Nativization and Endonormative Stabilization Phase in Schneider’s (2003) model, users of English in Outer Circle countries reinterpret and restructure complex subsystems of the input variety in subtle ways. This reorganisation proceeds without producing structures that in themselves are unacceptable in the historical input variety. (Huber 2012b: 240)

Being a well-studied area for the native standard varieties of English, the system of future time expressions (henceforth: FTEs) has received much less attention both in studies of second languages as well as in studies of New Englishes. One possible reason for this is the fact that the variation of WILL and BE GOING TO is rather one of tendencies than of rules in native English. The overuse of a certain variant does not generally produce errors – thus the study of FTEs has so far not been in the focus of researchers in the field of language contact. The only exceptions to this are purely quantitative in nature such as the studies by Collins (2009) and Collins & Yao (2012), who both compare frequencies of modals and semi-modals in different World Englishes, including WILL and BE GOING TO, as well as the studies by Sand (2005) and Deuber, Biewer, Hackert & Hilbert (2012), who both study quantitative differences in the use of FTEs between different New Englishes and BrE.

As corpus-based studies of native varieties have shown, the system of FTEs varies considerably across genre and style (cf. Mair & Leech, Mair 2015). Interestingly, the way in which the variation between the FTEs is constrained in spoken language seems to be relatively consistent across the major standard varieties of English, i.e. British and American English, where many of the differences between FTEs have been found to be of a lexical/syntactic and pragmatic nature (cf. Szmrecsanyi 2003; Berglund 2005; Torres Cacoullos & Walker 2009). It is thus expected that the present study will point to those differences between FTEs that are less readily acquired and transmitted in language contact situations.

This paper will focus on the variation between the modal WILL and its variants will, 'll and won’t and BE GOING TO with its variants be going to and be gonna. It is part of one of the studies of my PhD-thesis on tense, modality and aspect in Ghanaian English (Schneider 2015). The analysis is based on a corpus of 144,000 words of spoken private conversations of GhE. A set of comparable texts of BrE (direct conversations from the category S1A; ca. 180,000 words) from the International Corpus of English – Great Britain are used as a basis for comparison.
Using variationist methods such as mixed effects logistic regression models, I examine the differences between GhE and BrE in the future time system. It will be shown that while the two FTEs are not as clearly distributed syntactically and semantically in spoken GhE than in spoken BrE, the major differences between the two varieties in the FTE system cannot be found on the macro-level but on much lower levels, which corroborates previous findings on innovations in New Englishes (cf. Schneider 2007: 86–88).

The remainder of the paper is structured as follows: Section 2 will briefly capture previous research on future time marking in English and comment on some aspects of future time marking in indigenous Ghanaian languages. Section 3 will introduce the theoretical background of GhE and the data which will be used in the present study. In Section 4 I will give a description of how the FTEs have been counted, present an overview of the frequencies of the variant forms of FTEs in BrE and GhE and introduce the factors and the respective levels for which all instances of FTEs have been coded. Section 5 will present the results of three mixed effect logistic regression models on the variation of WILL and BE GOING TO in the two varieties and discuss the individual differences between them. Section 6 will then summarize and discuss the findings and conclude with some remarks on issues for future research.

2. Future time marking in English: Previous research

In an attempt to describe the differences between the various options to refer to the future in English most grammars focus on semantic (e.g. Quirk, Greenbaum, Leech, Svartvik & Crystal 1985; Leech 2004; Declerck 2006) or sociolinguistic (i.e. stylistic, regional) differences (e.g. Biber, Johansson, Leech, Conrad & Finegan 1999) between the different future time expressions (FTEs). WILL is the most frequent option and often considered the default variant. BE GOING TO is frequently attributed meanings such as ‘future fulfillment of the present’ (Fleischman 1982; Quirk et al. 1985; Leech 2004) or ‘determination’ (Palmer 1979). However, the claim that primarily semantic differences account for the choice of either WILL or BE GOING TO has been subject to many controversies. While there are certain environments in which the FTEs carry different meanings and “occupy lexical, syntactic, and pragmatic niches”, many differences are “largely neutralized in discourse” (Torres Cacoullos & Walker 2009: 321). While BE GOING TO is highly grammaticalized and becomes more and more frequent in those contexts in which it is interchangeable with WILL (cf. also Mair 1997 on written registers), WILL is more and more ‘pushed back’ into 1st person, affirmative, declarative, main clause contexts (cf. Szmrecsanyi 2003; Berglund 2005; Torres Cacoullos & Walker 2009;
Tagliamonte, Durham & Smith 2014). However, while most constraints seem to be of a lexical and syntactic nature, WILL persists in its original meaning of ‘willingness’ or ‘volition’ in highly specific contexts implying offers, requests and promises or threats, resulting in a number of semantic constraints associated with agentivity and adverbial modification (cf. also Torres Cacoullos & Walker 2009).

As mentioned above, little has been said about the variation of WILL and BE GOING TO in New Englishes. Sand (2005) reports lower numbers of BE GOING TO and higher numbers of WILL in a number of New Englishes. Deuber et al. (2012) confirm this finding in their data from Indian English, Singapore English and Fiji English, but also note that the Caribbean Englishes make more use of the BE GOING TO future – especially those that have substrates with FTEs based on motion verbs (e.g. Trinididian Creole, Bahamian Creole). Collins (2009) and Collins & Yao (2012) report fewer uses of semi-modals (including be going to, have to, need to, be able to etc.) in Outer Circle varieties than in Inner Circle varieties, which could point towards more conservative future and modality systems. For Indian English, Berglund (2005: 121) notes that, irrespective of text category, there are fewer tokens of the variant ‘ll and also fewer tokens of going to/gonna, but more attestations of will and shall than in BrE and American English (AmE).

For GhE, quantitative research on future time marking is as scarce as on the other New English varieties, although some sources suggest the development of a new FTE on the basis of the verb come – a construction formally similar to the BE GOING TO construction (cf. Huber 2012a). The present data showed only four instances of this construction (cf. Schneider 2015: 125–126). The distinction between future and non-future is the temporal opposition that is the most central and the most frequent in the Kwa languages, the largest group of indigenous Ghanaian languages. All languages of this group have at least one fully grammaticalized overt expression to refer to the future (Ameka & Kropp Dakubu 2008b). Akan, for example, has three to four different means to refer to the future. One of these is the future affix be, also referred to as the Future Aorist (Boadi 2008). According to Osam (2008) and Boadi (2008), this marker has derived from the verb be ‘to come’. Accordingly, it does not combine with the motion verbs in the language (cf. Osam 2008), pointing to the retention of its original meaning. In addition, the time marker na (meaning ‘then’) can be used to refer to future situations, once future time reference is established. This marker also combines with aspektual affixes (Progressive) and tonal changes (habitual and continuative) (cf. Osam 2008; Boadi 2008). As in English, the Progressive itself can be used to refer to future situations, once future time reference is established (Boadi 2008). In negative contexts, the difference between ‘future’ and ‘progressive’ meaning is neutralized (cf. Osam 2004). Furthermore, the combination between the motional affix be ‘to come’ and progressive has also been classified as ‘immediate future’ (cf. Dolphyne 1988).
Future markers based on movement verbs are quite widespread in the indigenous languages of Ghana, especially from source verbs meaning ‘come’. In Ghanaian Pidgin English the general future and hypothetical marker is *go* (Huber 2008b).

3. Ghanaian English: Theoretical background and data

In Ghana, as in many other African countries, the use and status of English as an official language is a result of colonial history. The language has gained a high number of (especially second language) speakers and has seen an extensive increase in usage since independence in 1957. English is the language for official government business in the executive, courts of law, parliament, civil service, in the national media, in the army and for preaching in many Orthodox Churches, especially in urban areas (Dolphyne 1995). It is also the medium of instruction across all levels of education and is in fact increasingly becoming a first language for some of the indigenous people. English as L1 is especially rare among urban lower classes and tradition-bound people in the countryside. It is rather a characteristic feature of the middle-to-upper class of the nation (cf. Huber 2008a). There are Ghanaianisms heard on all levels of language, but local innovations seem to be much more readily accepted on the levels of phonology and the lexicon than in the areas of grammar. With respect to its diachronic development, GhE thus falls within Stage 3 “Nativization” and Stage 4 “Endonormative Stabilization” in terms of Schneider’s (2003, 2007) Dynamic Model (cf. Huber 2014).

For the analysis of spoken GhE 60 private face-to-face conversations were included, 45 of which will eventually be part of the *International Corpus of English* (ICE) Ghana component (henceforth ICE-GH).¹ All data were collected in Ghana during fieldtrips in the years 2002, 2008 and 2010.² A total of 144,000 words of GhE were included in this study. I will refer to it here as CS-GH (Corpus of Spoken Ghanaian English).

CS-GH includes data from 133 different speakers. 80% of the speakers are university students in their early or mid 20s. Two thirds of them are male. Although

---

1. ICE-GH is a joint project by the University of Giessen, Germany, and the University of Ghana. It is still under compilation. By the time of analysis the texts were of variable length. For an overview of the text lengths see Schneider (2015).

2. The recordings were transcribed and annotated either by the present researcher, by members of the ICE-GH team or by visiting students from Ghana at the University of Giessen. Each transcript was double-checked by a Ghanaian native speaker as well as by the present researcher. I am grateful to Agoswin Musah for taking the time to check all transcripts that are included in the dataset.
this unequal distribution is the result of the methodology used in recruiting
speakers for the recordings, to some extent, it also reflects reality in the sense that
more and more young people complete senior secondary school and thus qualify
as speakers for ICE. The variety analyzed in the present work could thus be char-
acterized as ‘English as spoken by young, educated Ghanaians’. Although this does
not capture the complete picture of the speech community of speakers of GhE,
it probably gives a good impression of where GhE is heading towards, as young,
educated speakers are most likely to be those who influence the development of
the local standard. Irrespective of sex, age and occupation, all speakers included
in the data fulfill the general guidelines set by the ICE team in terms of minimum
education (Senior Secondary School must be completed), nationality (speakers/
authors have to be natives of the respective country (born or settled there at an
eyear age)) and acquisition of English (speakers need to have received their educa-
tion through the medium of English in the respective country).

With respect to the L1s of the speakers, the present dataset is quite a good
approximation of the distribution of L1s across the population according to the
Ethnologue (Lewis, Simons & Fennig 2014). Out of 40–80 different languages spo-
ken in Ghana a total of 19 different languages are represented by the speakers in
CS-GH. With more than 75%, the Kwa languages are slightly overrepresented in
the data as compared to estimated figures provided in the Ethnologue. The major
Gur languages, Dagbani and Dagaare, by contrast, are underrepresented, although
some other, minor languages from that branch (e.g. Gurune, Kusaal) make up for
the differences. For a detailed overview, see Schneider (2015).

4. Counting and coding the future

For the analysis of variation in future time marking all the future temporal refer-
ence constructions in the data were extracted and coded following the methodol-
ogy developed in Schneider (2015), based on the relevant literature on modality
such as Coates (1983), Palmer (1979, 1990, 2001), and Bybee, Pagliuca & Perkins
(1994), but also on descriptions of modality given in grammars such as Quirk et
al. (1985), Biber et al. (1999), Huddleston & Pullum (2002), Declerck (2006) and
Aarts (2011). I included only instances of WILL (including WILL + Progressive)
and BE GOING TO that were clearly temporal. SHALL was not considered giv-
en its low frequencies in both corpora. Futurate uses of the Present Progressive
were not included in the analysis of the present paper, but for an analysis see
Schneider (2015).3

3. In CS-GH there are 109 tokens of the futurate Progressive, spoken ICE-GB shows 199 tokens.
With regard to WILL, only those tokens were considered as FTEs that make reference to the future, and can be ascribed the function of expressing the meanings of ‘prediction’ and ‘intention’ or ‘indeterminate’. The following three examples illustrate these uses.

1. **Prediction**: But I’m sure the lights will come back soon and you can study something. (GHSM_0472, CS-GH S1A-041)

2. **Intention**: I will name the first one Mercedes the second one Champagne the third one Paula and the boy Gerald. (GBSM_0322, CS-GH S1A-022)

3. **Indeterminate**: You call yourself a university lady that somebody will marry some day. (GHSM_0367, CS-GH S1A-032)

Meanings of WILL that do not refer to the future include the meanings of habit, present or habitual predictability and willingness, although a number of them are mergers between various meanings. Those instances of WILL which seemed to be mergers between predictability and prediction were eventually categorized as cases of prediction (with a rather strong epistemic meaning) and assigned FTE status. The following example illustrates such a case.

4. That’s goat jollof. Try that one and see. You will like it I tell you. (CS-GH S1A-010)

Mergers between willingness and intention were categorized as instances of intention, as in (5), whereas highly pragmatic uses of WILL for making offers or requests such as (6) were categorized as instances of willingness, and thus not counted as FTEs.

5. Because me when he I’ve I’ve been telling him that when he he backs out I won't enter into a relationship again. (CS-GH S1A-016)

6. Dad will you have some more juice? (ICE-GB S1A-022)

All variant forms of WILL (will, ’ll, won’t) were subsumed under WILL, although it has been suggested that ’ll and will should be considered independent future forms because they are found in quite diverse patterns (e.g. Berglund 2005; Nesselhauf 2010). However, as Torres Cacoullos & Walker (2009: 340) argue, ’ll and will should rather be seen as variants in complementary distribution, and the favoring of the
contracted form by the first person could well be interpreted as a frequency effect. While ‘ll is most pervasive with the meaning of ‘intention’, this is due to its strong association with the 1st person singular and less to a distinct meaning of the form.

Some accounts treat WILL + Progressive separately from WILL + infinitive because the former construction has been described as developing into an independent future construction (cf. Celle & Smith 2010). WILL + Progressive exhibits a wide variety of functions. Only in examples like (7) below are we dealing with cases in which the construction WILL + Progressive is the sum of its components, ‘future’ and ‘ongoingness’.

(7) You’ll be sitting on the phone at work ringing around. (GBSM_0914, ICE-GB S1A-079)

In other cases the use of the WILL + Progressive-construction applies to a single event viewed in its entirety, something that is commonly referred to as ‘future-as-a-matter-of-course’ (cf. Leech 2004; Aarts 2011: 286) i.e. a future situation the speaker is certain about, as in example (8).

(8) The SRC elections will be coming off in two weeks. (CS-GH S1A-019)

Finally, the use of the WILL + Progressive-construction is used in questions for purposes of disambiguation. While the use of WILL + infinitive often receives the reading of an invitation, request or offer, the use of WILL + Progressive avoids such an interpretation (cf. Declerck 2006: 343), as in example (9).

(9) Will you be reading the Bible this Sunday at church? (CS-GH S1A-046)

As mentioned by various authors (e.g. Coates 1983; Declerck 2006), with WILL + Progressive the reference is mostly to ‘pure future’ or ‘prediction’, i.e. without any implication of willingness or intention. However, as noted by Celle & Smith (2010) there are examples in which expressions with this construction could equally receive a volitional reading. Although its disambiguating power, as well as its strong deterministic meaning component, seem to be important features of the construction WILL + Progressive, in other contexts its use seems to be merely a more tentative or more colloquial way of referring to the future. While the use of this construction deserves more detailed analysis, its uses in the data were too infrequent to make any generalizations. As noted in Schneider (2015) on the Progressive, the frequent use of modal progressives in GhE may well be a purely formal choice rather than motivated semantically or pragmatically. It was thus decided not to analyze instances of WILL + Progressive as independent FTEs but to lump them together with instances of WILL + infinitive.

With respect to BE GOING TO, the same regulations for FTE assignment were applied as for the modal WILL: All cases of BE GOING TO that could be
analyzed as cases of ‘intention’, ‘prediction’ or ‘indeterminate’ were categorized as FTEs.\(^5\) Examples of these meanings are given in (10)–(12).

(10) **Intention:** I’m not going to talk in this room again. I’m going to communicate with you girls through sign language. (CS-GH S1A-046)

(11) **Prediction:** He is going to die. (CS-GH S1A-003)

(12) **Indeterminate:** And nobody is going to referee them, too. (CS-GH S1A-009)

In very few cases in BrE, BE GOING TO has the meaning of ‘present predictability’ rather than ‘prediction’. If reference was not clearly to the future, as in (13) and (14), these occurrences were not considered.

(13) This is going to be a question of who you know not what you know. (ICE-GB S1A-027)

(14) Even though we’ve got this wretched document we’re talking about there’s always going to be <,> an Asterix book by the bedside <,> or something like that. (ICE-GB S1A-013)

All those cases of BE GOING TO that were clearly spatial in meaning were deleted. Instances of past forms of BE GOING TO as in What was I going to say were excluded from the analysis as they are either analyzed as having the meaning of ‘past intention’ (either as ‘unfulfilled situation’), or as a marker of reported speech as in he said they were not going to do the …. In this respect they are not future or modality markers but purely aspectual in nature.

In CS-GH only 537 out of 916 tokens of WILL (58.65%) qualify as instances of FTEs, whereas in spoken ICE-GB 478 out of 623 tokens of WILL (78%) qualify as FTEs. As shown in Schneider (2015), the modal WILL shows a considerably higher frequency in spoken GhE than in spoken BrE, both in absolute as well as in relative numbers, which is not only due to a higher use of WILL as a FTE in GhE but also to other differences in the use of the modal. The first one is the use of WILL instead of WOULD in GhE, either to refer to hypothetical situations or in those contexts in which WOULD would yield the effect of ‘tentativeness’ or ‘politeness’. The second one is the extensive use of habitual WILL, WILL being frequently used as a marker of habitual behavior and typical activities in this variety. These usage patterns show considerable parallels to some of the indigenous Ghanaian

\(^5\) I applied the same conventions and practices in categorizing instances of BE GOING TO as either ‘intention’, ‘prediction’ or ‘indeterminate’ as was done for the modal WILL (see footnote 4). The categorization was mainly based on grammatical person, agentivity of verb, sentence type and, for some cases, discourse cues.
languages, as, for example, to the potential morpheme *la* in Ewe (Essegbey, email conversation). For a comparison of the use of WILL and BE GOING TO in native and non-native varieties of English the precise use of a token has to be taken into consideration in order to avoid premature conclusions on the advancement of a variety in terms of the use of semi-modals.

Figure 1 gives an overview of the frequencies of FTEs per 100,000 words in CS-GH and spoken ICE-GB. Additionally, Table A.1 in the appendix gives raw frequencies. For the sake of explicitness, the figure shows WILL and WILL + Progressive separately. WILL is further divided into *will*, *’ll* and *won’t*. In Table A.1 I further distinguish between *be going to* and the phonologically reduced form *be gonna*, although it has to be kept in mind that the transcription conventions for spoken texts in the corpora might actually obscure the findings.

![Figure 1](image.png)

**Figure 1.** FTEs in CS-GH and Spoken ICE-GB: Normalized Frequencies of Variant Forms

Remarkable is the high frequency of the full form *will* in CS-GH and the relatively low frequency of the contracted form *’ll*. In spoken ICE-GB the picture is reversed, and *’ll* is more than twice as frequent as the full form.\(^6\) *be going to/be gonna* and *’ll* are thus the most characteristic FTE forms of spoken BrE. On the one hand, these differences might show the less informal character of spoken GhE in comparison to BrE. On the other hand, it could be both the result of and the reason for lower uses of the collocation with the 1st person singular *I’ll*. I will turn to these issues again below.

---

\(^6\) This difference is highly significant with \(p < 0.001\).
Table 1 gives an overview of absolute and relative frequencies of WILL and BE GOING TO in the two corpora.

Table 1. FTEs in Spoken ICE-GB and CS-GH: Relative and Absolute Frequencies

<table>
<thead>
<tr>
<th></th>
<th>Spoken ICE-GB</th>
<th>CS-GH</th>
</tr>
</thead>
<tbody>
<tr>
<td>WILL</td>
<td>63.95% (488)</td>
<td>81.36% (537)</td>
</tr>
<tr>
<td>BE GOING TO</td>
<td>36.04% (275)</td>
<td>18.63% (123)</td>
</tr>
</tbody>
</table>

As the above table suggests, GhE makes much more use of the modal WILL to refer to the future. The use of BE GOING TO is much less frequent. Judging from the raw frequencies one might assume that GhE is simply less advanced in the use of the semi-modal BE GOING TO. In the next section I will show that there are not only differences between the two corpora in terms of frequencies of FTEs but also in terms of patterns of use.

Each token of WILL and BE GOING TO which qualified as an instance of a FTE as explained above was coded for a number of factors based on the findings in the literature. The factors included here were selected on the basis that they are measurable and not intuitive. For example, WILL and BE GOING TO have been described to differ on semantic grounds, as, for example with respect to modality type (prediction, intention, etc.), degree of certainty, conditionality (cf. Declerck 2006: 352–355; Leech 2004: 59), or determination (Palmer 1979). Semantic factors were either determined on the basis of the context (temporal adverbial modification), or were excluded from the analysis (conditionality, modality type, degree of certainty, determination).

4.1 Sentence type

Torres Cacoullos & Walker (2009) report sentence type as the most important factor in the choice of FTEs in their data of Quebec and Montreal English. According to Tagliamonte et al. (2014), 1st person singular interrogatives such as What am I going to do? are the locus of development of the BE GOING TO future. Declerck (2006) mentions the preference for BE GOING TO, the Progressive and WILL + Progressive in interrogatives as WILL + infinitive in interrogatives often gets the default volitional reading of ‘request’, ‘offer’, or ‘invitation’. Hence, I categorized all tokens in the sample according to whether they occurred in interrogative (yes/no-questions and wh-questions) or declarative sentences.
4.2 Polarity

Several studies have identified polarity as an important predictor for the choice of FTEs in English (cf. Berglund 2005; Szmrecsanyi 2003; Tagliamonte 2013; Tagliamonte et al. 2014; Torres Cacoullos & Walker 2009). According to these studies, BE GOING TO has a much higher share of tokens in negated contexts than in affirmative contexts.

Explanations for the choice of BE GOING TO rather than WILL in negative contexts commonly include the possibility for negated WILL to receive a reading of ‘refusal’ when occurring with agentive verbs (cf. Declerck 2006; Coates 1983). Szmrecsanyi (2003), however, also notes regional differences with respect to negated contexts and choice of FTE: While American English prefers not be going to or not gonna, BrE favors the use of contracted won’t. Apparently, the choice thus seems to be led by preferences for forms rather than semantics. In this study I will only consider those sentences or clauses as negative which are negated by not (including variants like won’t).

4.3 Subject type

Subject type has been reported to be an extremely important factor in the choice of FTEs in several varieties of English (Berglund 2005; Poplack & Tagliamonte 1993; Tagliamonte et al. 2014; Tagliamonte 2002; Torres Cacoullos & Walker 2009; Wekker 1976). 1st person subjects seem to be linked to the use of WILL, whereas 2nd and 3rd person subjects are associated with the use of BE GOING TO. Torres Cacoullos & Walker (2009) admit that a correlation between 2nd person subjects and interrogatives is likely to account for the use of BE GOING TO. Nevertheless, the use of 1st person subjects and WILL seems to be due to a strong collocational attraction of the pronoun to the contracted form of WILL. For the present analysis I set up two different levels for the factor subject type: 1st person singular/plural subjects, and non-1st person subjects (2nd person singular/plural, 3rd person singular/plural animate and 3rd person singular/plural inanimate subjects).

4.4 Temporal adverbial specification

Temporal adverbial specification is a frequently cited criterion for the choice of FTEs. As Leech notes, “a sentence with will describing a future event feels incomplete without an adverbial of definite time” (2004: 57), and uses without any temporal adverbial specification may sound odd in some cases, as in She will have twins. In contrast, BE GOING TO is said to be perfectly acceptable without any temporal specification, as this construction “expresses dual time reference with
emphases on the present” (Declerck 2006: 346), i.e. has a temporal anchor in the present. Declerck (2006: 351) explains that this is because it represents the post-present (i.e. future) actualization of a situation as related to the present. He adds that “[b]ecause of this immediacy implicature, be going to can be used without an adverbial or contextual specification of a specific future time” (2006: 351). According to Coates:

> [A] crucial feature of the meaning of BE GOING TO, both Root and Epistemic, is that the future event or state referred to in the main predication is seen as happening very soon after the moment of speaking and as being related to the present.

(1983: 198)

As Leech (2004: 59) points out, “when the clause with be going to contains no time adverbial, immediate future is almost certainly implied […] unless some adverbial indicates otherwise” – hence the association of BE GOING TO with near future situations. Indefinite futures, it has been shown, are associated with WILL, rather than with BE GOING TO (Torres Cacoullos & Walker 2009). In order to test these hypotheses on the basis of my data, I set up three different categories: (1) no temporal adverbial specification, (2) indefinite time adverbials such as never, always, later etc., (3) definite time adverbials.

4.5 Agentivity

Agentivity is crucial in the development of FTEs in general. In early stages of grammaticalization, both WILL and BE GOING TO are restricted to agentive verbs, as their original meanings are ‘volition’ and ‘movement’, respectively (cf. Bybee et al. 1994). In some contexts, these meanings are still inherent. I distinguish two levels, agentive and non-agentive verbs. Agentivity had to be carefully assigned to each individual verb, as this class is highly context-dependent. For example, (to) learn is actually not agentive; however, in GhE it often has the meaning of ‘to study’, and then qualifies as agentive.

5. An analysis of WILL vs. BE GOING TO

For the analysis of variation between WILL and BE GOING TO a multivariate analysis on the basis of mixed effect logistic regression7 was carried out. Logistic

---

7. A regression model is a statistical process for estimating the relationship between factors that are assumed to play a role in the distribution of variants. Logistic regression is a model where the variable of interest is categorical.
regression has a long-standing tradition in quantitative, variationist sociolinguistics as it predicts the joint effect of intralinguistic and/or extralinguistic factors on the choice of a binary dependent variable. While conventional logistic regression models only predict the impact of fixed effects, i.e. factors that are repeatable over datasets, mixed effects models also consider the impact of random effects, i.e. that have infinitely many levels that are not predictable nor repeatable (cf. Baayen 2008: 263). For the present study, I included both speaker and verb as random effects. This has two major advantages: First, including speaker as a random effect will compensate for the imbalance in the amount of FTE tokens among the different discourse participants, so that no single speaker influences the results of the study too extremely. Secondly, the random effect for speaker can itself be used as a measure for how extremely individual speakers differ from the mean. Similarly, taking verb as a random effect both compensates for the differences in representation of individual verbs and gives insight into the behavior of individual verbs.

I included five main effect predictors (sentence type, subject type, polarity, agentivity, temporal adverbial modification). In order to obtain a clear picture of where the differences lie in terms of constraints between the two varieties, I first fitted a model for both varieties which captures the differences via interaction effects. This has the advantage that interaction effects can assess the significance of variety differences. In a second step I fitted a model for each variety separately as this will facilitate the interpretation of the interaction effects between the varieties. I generated the models including all fixed effect main predictors, all interactions as well as random effects for speaker and for verb, using the function glmer() from the package lme4 in R (Version R 3.2.2 GUI 1.66 Mavericks build). By stepwise deletion of insignificant main predictors and interactions, I then arrived at the minimal adequate model (for the procedure cf. e.g. Gries 2009). In order to avoid collinearity, I calculated the generalized variance inflation factor for each of the predictors in both models.

Table 2 reports fixed and random effects in the minimal adequate mixed effect logistic regression model, Model 1, for variety differences between BrE and GhE. In addition to the five main effect predictors, I included interactions between

---

8. R is a programming language which is also increasingly used for statistical computing in linguistics.

9. Collinearity occurs if the predictors in a regression model enter into strong correlations. This is problematic because it makes it difficult to ascertain which of the predictors has explanatory value (Baayen 2008: 198).

10. The generalized variance inflation factor (gvif) was calculated using the function vif() from the package car in R. As none of the gvifs for particular predictors was above 10, all main predictors and interactions were maintained in the model. For descriptions of the gvif, cf. Fox (1997).
another predictor, variety, and all other predictors. Moreover, I included random
effects for both verb and speaker.\textsuperscript{11} All predicted estimates are for the modal WILL.
The model has a good fit (C = 0.8883, Somers’ Dxy = 0.7766) and predicts 82.24% of all outcomes correctly, which is 9.91% above baseline (baseline: 72.33%).

Table 2. Results of Minimal Adequate Mixed Effect Logistic Regression Model 1: CS-GH and Spoken ICE-GB

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>(Intercept)</td>
<td>0.2991</td>
<td>0.5469</td>
</tr>
<tr>
<td>Speaker</td>
<td>(Intercept)</td>
<td>1.2974</td>
<td>1.1391</td>
</tr>
</tbody>
</table>

Number of observations: 1442, Groups: Verb, 269; Speaker, 265

Fixed Effects:

|                  | Estimate | Std. Error | z value | Pr(>|z|) |
|------------------|----------|------------|---------|---------|
| (Intercept)      | 1.3899   | 0.2134     | 6.513   | 7.37e-11 *** |
| Variety          |          |            |         |         |
| CS-GH            | 0.9361   | 0.3397     | 2.755   | 0.005865 ** |
| Sentence type    |          |            |         |         |
| Interrogative    | −1.5297  | 0.3082     | −4.964  | 6.92e-07 *** |
| Subject type     |          |            |         |         |
| Non-1st person   | −1.3016  | 0.2366     | −5.502  | 3.76e-08 *** |
| Polarity         |          |            |         |         |
| Negative         | −0.4848  | 0.2247     | −2.157  | 0.030983 * |
| Agentivity       |          |            |         |         |
| Non-agentive     | 0.8770   | 0.2554     | 3.434   | 0.000595 *** |
| Temporal adverbial specification | | | | |
| Definite         | −0.3562  | 0.2311     | −1.541  | 0.123212 |
| Indefinite       | 0.8486   | 0.3833     | 2.214   | 0.026847 * |
| Variety*Subject type | | | | |
| CS-GH*Non-1st person | 0.9899  | 0.3713     | 2.666   | 0.007667 ** |
| Variety*Agentivity | | | | |
| CS-GH*Non-agentive | −0.8650 | 0.3551     | −2.436  | 0.014856 * |

\textsuperscript{11} Call: GBGH.glmer <- glmer(GBGH$FTE2 ~ Variety*(SenT + Pol + SubT + Adv + Agentivity) + (1|Speaker) + (1|Verb), data=GBGH, family=binomial, control=glmerControl(optCtrl=list (maxfun=100000))).
Let us discuss the results of Table 2 step by step. The top lines in the table report the standard deviations for the random effects for verb and speaker. For random effects for verb, for example, each individual verb receives its own adjusted intercept, which is either positive or negative. If a particular verb receives a negative adjusted intercept, this means that there is a preference for *BE GOING TO* for this verb after considering the predictions of the fixed effects that the model makes for the choice of FTEs. The higher the adjusted intercept departs from 0, the larger the effect size for the choice of a particular verb for a FTE. For high-frequency verbs, a high adjusted intercept could point to a strong collocation between a verb and a particular FTE. The random effects for verbs and speakers will be considered in more detail below. The line below the values for random effects gives an overview of the number of observations considered in the model (1442), the number of individual verbs considered (269) and the number of speakers included (265).

The subsequent lines give an overview of the significant fixed effect predictors in the model. The figure reports the coefficient estimates, which are expressed in logits. As stated above, predicted estimates are for the modal *WILL*. Concerning random effects, a positive coefficient indicates that the probability for the occurrence of *WILL* increases, while a negative value indicates that the probability for *WILL* decreases (and thus increases for *BE GOING TO*). The higher the value of a coefficient estimate, the stronger the effect size of a predictor. Each coefficient estimate is further accompanied by its standard error, a z-value and the significance level. Now consider the predictor variety: The default level of this factor is ‘ICE-GB’. The positive coefficient estimate for the level ‘CS-GH’ (−1.5297) means that speakers in CS-GH contexts are favor WILL more than speakers of ICE-GB do. As the p-value indicates, this predictor is very significant (*p* < 0.01).

Sentence type, subject type, polarity, temporal adverbial modification and agentivity are all significant in Model 1. Furthermore, variety is a significant predictor for FTE choice. Turning to the most important aspect of this model, the interactions between variety and other main effect predictors, we observe that there are two significant interactions between variety and other main predictors. There is a very significant interaction between variety and subject type (*p* < 0.01), and a significant interaction between variety and agentivity (*p* < 0.05). The differences in future marking between the two varieties thus lie within the domains of agentivity and subject type. In order to understand the differences between the two varieties, we will now have a look at the models for each of the corpora individually. I will start with the model for spoken ICE-GB.
Table 3 shows the results for the minimal adequate mixed effect logistic regression model for spoken ICE-GB, Model 2.\textsuperscript{12} Model 2 considers 760 datapoints from 149 speakers. The model considers 159 different verbs. It has a good fit ($C = 0.8437$, Somers’ $D_{xy} = 0.6875$) and predicts 77.23% of all outcomes correctly, which is 13.28% above baseline (baseline: 63.95%). The standard deviation for random effects for verb is at 0.5633, and the standard deviation for random effects for speaker is at 0.7039.

Table 3. Results of Mixed Effect Logistic Regression Model 2: Spoken ICE-GB

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>(Intercept)</td>
<td>0.3173</td>
<td>0.5633</td>
</tr>
<tr>
<td>Speaker</td>
<td>(Intercept)</td>
<td>0.4955</td>
<td>0.7039</td>
</tr>
</tbody>
</table>

Number of observations: 760, Groups: Verb, 159; Speaker, 149

| Fixed Effects | Estimate | Std. Error | $z$ value | $Pr(|z|)$ |
|---------------|----------|------------|-----------|----------|
| (Intercept)   | 1.3272   | 0.1986     | 6.682     | 2.36e-11 *** |
| Sentence type |          |            |           |          |
| Interrogative | −1.3257  | 0.3679     | −3.604    | 0.000314 *** |
| Subject type  |          |            |           |          |
| Non-1st person | −1.2222  | 0.2301     | −5.312    | 1.08e-07 *** |
| Polarity      |          |            |           |          |
| Negative      | −0.6207  | 0.2717     | −2.284    | 0.022354 *  |
| Agentivity    |          |            |           |          |
| Non-agentive  | 0.7277   | 0.2545     | 2.860     | 0.004243 ** |
| Temporal adverbial specification | | | | |
| Definite      | −0.3618  | 0.2649     | −1.366    | 0.172093 |
| Indefinite    | 1.4676   | 0.6682     | 2.196     | 0.028078 * |

As Table 3 shows, temporal adverbial specification, sentence type, subject type, agentivity and polarity – starting from the predictor with the largest effect size (see coefficient estimates) to the one with the smallest – are all significant predictors for the choice of FTEs in ICE-GB.

\textsuperscript{12} Call: \texttt{LRGB.glmer <- glmer(LRGB$FTE2~SenT + Agentivity + Pol + SubT + Adv + (1|Speaker) + (1|Verb), data=LRGB, family=binomial, control=glmerControl(optCtrl=list (maxfun=100000))}).
Figure 2 visualizes the effect size of the significant fixed effect predictor levels on the choice of the FTEs. Columns stretching to the right are predictor levels favored by WILL, whereas columns stretching to the left are predictor levels favored by BE GOING TO.

Table 4 shows the mixed effect logistic regression model for CS-GH, Model 3. Model 3 considers 682 data points from 52 files and 116 speakers. The model considers 182 different verbs. The model has a very good fit ($C = 0.9511$, Somers’ $Dxy = 0.9023$) and predicts 88.12% of all outcomes correctly, which is 6.36% above baseline (baseline: 81.76%). The standard deviation for speaker is very high (1.7202), indicating that there is a high variability among the speakers within the variety. The standard deviation for verb is also higher than in Model 2 for spoken ICE-GB (0.9782).

As Table 4 shows, only sentence type is a significant predictor for the choice of FTEs. As in ICE-GB, BE GOING TO favors interrogative environments (see the negative coefficient estimate for interrogatives in the table). In contrast to ICE-GB, subject type, agentivity, temporal adverbial modification and polarity do not play an important role in the choice of FTEs in CS-GH. There are thus far fewer fixed effect predictors that govern the use of FTEs in CS-GH than there are in ICE-GB.

13. Call: LRGH.glmer <- glmer(LRGH$FTE2~SenT*Agentivity + Pol*Agentivity + CIT + SubT + Adv + (1|Speaker) + (1|Verb), data=LRGH, family=binomial, control=glmerControl(optCtrl=list(maxfun=100000))).
Table 4. Results of Mixed Effect Logistic Regression Model 3: CS-GH

**Random Effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>(Intercept)</td>
<td>0.957</td>
<td>0.9782</td>
</tr>
<tr>
<td>Speaker</td>
<td>(Intercept)</td>
<td>2.959</td>
<td>1.7202</td>
</tr>
</tbody>
</table>

Number of observations: 682, Groups: Verb, 182; Speaker, 116

**Fixed Effects:**

| Sentence type | Estimate | Std. Error | z. value | Pr(>|z|) |
|---------------|----------|------------|----------|---------|
| Interrogative | −2.2090  | 0.6030     | −3.664   | 0.000249*** |

Let us put the results of the three regression models into context: As was shown, in both varieties BE GOING TO favors interrogative contexts over declarative ones. This is the only environment for which the occurrence of BE GOING TO is predicted in GhE. The situation in BrE is very different: BE GOING TO is not only much more frequent in BrE than in GhE, it also ‘pushes’ WILL back into specific syntactic and semantic slots. For GhE the regression model did not identify such characteristic slots. Rather, it seems like the division of tasks between WILL and BE GOING TO which we can find in BrE is neutralized in GhE. However, with the exception of subject type, in CS-GH the two FTEs still favor the same linguistic environments as in spoken BrE, showing that most of the differences are quantitative rather than qualitative, with the interactions of variety with polarity and temporal adverbial modification in Model 1 not even being significant.

First, considering subject type: Subject type is only a significant predictor in ICE-GB. In ICE-GB more than 90% of all cases in which 1st person singular subjects occur with WILL in affirmative contexts consist of the contracted ‘ll forms, and 80% of all 1st person plural subjects with WILL are also found with the contracted form. By contrast, in CS-GH only 50% of all cases of 1st person singular subjects and only 57% of all cases of 1st person plural subjects occurring with WILL are cases of the contracted form. This points to the strong collocational force between the pronoun and the modal verb form in spoken BrE, which is not perceptible to this extent in spoken GhE. In addition, the number of 1st person singular subjects with WILL in ICE-GB is about 50% higher than in CS-GH, whereas the numbers for all other subject types occurring with WILL are lower in ICE-GB than in CS-GH. Apart from the distributional difference between the varieties there is thus also a discourse difference: Apparently, the speakers in CS-GH do not talk about their own plans, intentions, predictions etc. as much as speakers in spoken ICE-GB do (cf. also Schneider 2015).
The normalized frequencies of the FTEs by subject type are shown in Figures 3 and 4.

In contrast to subject type, for all other predictors WILL and BE GOING TO show the same preferences in CS-GH and spoken ICE-GB, but these preferences are only significant for ICE-GB. For example, polarity is an important fixed effect predictor in ICE-GB only. Negative contexts are associated with BE GOING TO

Figure 3. Normalized Frequencies of FTEs by Subject Type in Spoken ICE-GB

Figure 4. Normalized Frequencies of FTEs by Subject Type in CS-GH
more than affirmative contexts in both varieties, whereas in affirmative contexts WILL is preferred. The difference in percentages of the two FTEs between affirmative and negative contexts is larger in ICE-GB than in CS-GH. In ICE-GB, BE GOING TO reaches 50% in negative contexts, whereas in CS-GH it is still only at about 30%. This is shown in Figures 6 and 7.

The distribution of affirmative and negative contexts is very similar in the two varieties with a ratio of 9:1. All tokens of WILL in negated contexts in ICE-GB are instances of won’t, with only one exception. This indicates a strong use of the contracted form won’t in BrE, a tendency already mentioned by Szmrecsanyi (2003).

Figure 5. Distribution of FTEs in Spoken ICE-GB: Polarity

Figure 6. Distribution of FTEs in CS-GH: Polarity
In CS-GH, 12 out of 44 negated tokens of WILL are not instances of won’t. In negative contexts the collocated forms I’ll and we’ll are generally not used (there are four such cases in the CS-GH).

Agentivity was shown to be significant in spoken ICE-GB only, although BE GOING TO favors agentive contexts over non-agentive ones in both varieties. However, there are some interesting differences between spoken BrE and spoken GhE for negative and interrogative contexts: As Figure 7 shows, in ICE-GB, in

![Figure 7. Distribution of Forms in Negative Contexts by Agentivity in Spoken ICE-GB](image)

![Figure 8. Distribution of Forms in Negative Contexts by Agentivity in CS-GH](image)
negated sentences BE GOING TO is specifically preferred in agentive contexts, meaning that *not going to/not gonna* is favored in environments that express ‘no intention’ on behalf of the subject. This is different in CS-GH, as Figure 8 shows. Here, WILL is even more frequent with agentive verbs, whereas BE GOING TO is more frequent with non-agentive verbs. The distribution of WILL and BE GOING TO is not significant in GhE, though.

Looking at negated contexts in the data, it can be observed that especially in 1st person singular contexts there is a clear preference for the use of BE GOING TO in BrE (examples (15)-(16)), whereas quite a number of cases of *won’t* are found in GhE (example (17)).

(15)  *Well I am not going to do* all the furniture rubbish myself (ICE-GB S1A-030)

(16)  *I’m not just going to go and do* it off the top of my head (ICE-GB S1A-082)

(17)  *Me I won’t waste* my money to go and buy ticket and then go and watch beauty contests (CS-GH S1A-021)

Thus, it is possible to identify a division of tasks between individual variants in BrE: While in affirmative contexts 1st person (singular) subjects most likely occur with *’ll*, the use of *won’t* is dispreferred with agentive verbs, especially *I won’t* + agentive verb. This seems to be an environment in which WILL retains its original meaning of ‘willingness’ (or ‘refusal’) in BrE and in which BE GOING TO is preferred. In GhE, such a division of tasks cannot be identified on the basis of the present data. *Won’t* and *will not* are frequently used with agentive verbs in CS-GH, both with 1st person as well as with non-1st person subjects, with the meaning of ‘intention’ rather than ‘refusal’. An example of WILL in conjunction with the 1st person singular from CS-GH is given in (18).

(18)  *That’s why I will not* wait for maybe twelve thirteen years (CS-GH S1A-032)

Let us now consider agentivity in interrogative contexts. A look at the data shows that, similar to the observations made above, there are quite a few instances in which *won’t* is used with agentive verbs in interrogative contexts in CS-GH. Examples (19)–(21) illustrate this use.

(19)  *Now you no go you no won’t you go school again?* (CS-GH S1A-007)

(20)  *Won’t they involve him?* (CS-GH S1A-012)

(21)  *Won’t she collect money for it?* (CS-GH S1A-014)

In spoken ICE-GB, *won’t* rarely ever occurs in interrogative contexts. There are three instances of such uses, but these contain non-agentive verbs. Combinations
like *I won’t* + agentive verb or *won’t you/she/he* + agentive verb thus seem to be reserved for interpretations subsumed under ‘willingness’ (refusal or offer/threat/request) but are dispreferred in more general future contexts. In specific syntactic-semantic environments, *BE GOING TO* is the natural choice in BrE. There are thus clearer form-meaning correspondences in BrE than in GhE. The special semantic-pragmatic nuances that *WILL* still has in these environments in BrE seem to be obsolete in GhE.

Turning to temporal adverbial specification: It is not clear at this point whether temporal adverbial modification should be viewed as a semantic or a lexical/syntactic factor or a mixture of both. In spoken ICE-GB *WILL* favors indefinite temporal adverbials. Given that these often occur preverbally (e.g. *never, always, etc.*), the reason might lie within syntax rather than semantics. This would, however, also mean that adverbials such as *certainly*, which occur preverbally as well, should be favored by *WILL*. Subsequent research will have to shed more light on this issue. A look at CS-GH reveals that indefinite temporal adverbials also occur primarily with *WILL*, but the distribution of the two FTEs across the three factor levels is not significant.

Let us now look at the random effects in the models. Considering the individual verbs that have an exceptionally high or low adjusted intercept for random effects in ICE-GB, we can see that the most extreme cases are actually the high-frequency verbs *to come, to tell, and to be*. The first two are strongly associated with the use of *WILL*, whereas the latter is associated with the use of *BE GOING TO*. Since non-agentive verbs are associated more with *WILL* in the model (even if not significantly), there are a number of high-frequency stative verbs, like *(to) be* and *(to) have*, which collocate with *BE GOING TO* and thus have negative intercepts. As Hilpert (2008) notes, collocations with high frequency stative verbs such as *(to) be* are indicative of later stages of the grammaticalization of *BE GOING TO*. Similarly, a number of high-frequency agentive verbs, e.g. *(to) ask, (to) talk, (to) give, (to) tell, (to) get, and (to) come*, collocate with *WILL*. This points to some important collocations of verbs and FTEs in BrE. The corpus data shows that some of the examples containing the high-frequency verbs favoring *WILL* are fixed expressions, as in the following example.

(22) No. Glass is very expensive, *I’ll tell* you. (ICE-GB S1A-007)

A much lower frequency of confirmatory uses of constructions of the *I’ll tell you-* type in CS-GH might be one of the reasons for lower numbers of 1st person uses in GhE.

A look at the intercepts for the individual verbs in CS-GH shows that the high standard deviation of random effects for verbs is due to high intercepts for a number of verbs. These include high-frequency verbs such as *(to) do, (to) ask, (to) talk*
and (to) give, which are preferred by BE GOING TO. The verb (to) go is the only high-frequency verb which strongly collocates with WILL. Interestingly, apart from two exceptions, motion verbs, including go, come, leave and send (‘bring’ in GhE), are not associated with BE GOING TO. As suggested by Torres Cacoullos & Walker (2009), the dispreference of BE GOING TO for motion verbs could be attributed to the retention of the meaning of ‘movement’ in the construction. However, the meaning of ‘movement’ inherent in the meaning of the BE GOING TO construction could well be a feature of GhE proper. In fact, the meaning of GhE I’m going to come is actually ‘I will go and then come back’ (Musah, p.c.), which has parallels in some of the local languages (e.g. Akan or Hausa; cf. Boadi 2008, Osam 2008). As already mentioned in Section 2, the Akan future marker be, which is itself derived from the verb ‘to come’, does usually not co-occur with motional affixes. This could thus be seen as an environment in which a routinized expression from GhE proper, which is itself directly copied from substrate languages, puts constraints on the use of FTEs.

In contrast to Model 2 for ICE-GB, the intercepts for individual speakers vary considerably in Model 3 for CS-GH. We find values ranging from −4.04 to +1.67, which points to the heterogeneity in the choice of FTEs within the variety, i.e. that there are great qualitative and/or quantitative differences in the use of WILL and BE GOING TO between individual speakers. Twelve speakers exceed the threshold of the standard deviation, i.e. +/−1.6. All but one of them have negative intercepts, indicating that they tend to use BE GOING TO to a greater extent than the mean predicted by the model. Unlike for spoken ICE-GB, a great deal of variation in the use of BE GOING TO vs. WILL cannot be accounted for by fixed effect predictors (i.e. syntactic, morphological and semantic predictors) in CS-GH but is to a large degree dependent on random effects for speakers. In GhE the choice of BE GOING TO seems to be more dependent on individual speakers, as the random effects have a large standard deviation in the regression model.

Let us consider the use of WILL and BE GOING TO in discourse from one selected conversation. Especially, specific speakers make frequent use of the BE GOING TO-construction. It is interesting to observe that, at times, this even spreads to the other speakers in the conversation. The following excerpt is taken from the file with both the highest overall frequency of FTEs and the highest number of tokens of BE GOING TO. Whereas two of the speakers (speaker <$B> and speaker <$D>), almost exclusively use BE GOING TO, speaker <$A> sticks

14. According to Sey (1973: 2) in ‘broken English’, a kind of English typically associated with uneducated speakers, I go come means ‘I am going away, but I’ll be back’. This expression can be seen as a direct copy from a number of indigenous Ghanaian languages.
to WILL in most cases. Only in some instances do the speakers change the FTE, namely in those contexts in which a different FTE has preceded.

(23) <$D$> You see this time he is not got the crowd behind him the crowd is against him though they will accord him some respect but then they are not going to cheer him up like how they used to do when those days you see

<$A$> yes yeah

<$D$> well uhm with me I don’t think the crowd is going be a problem

<$D$> I think Mourinho is someone who is cool headed all the time he doesn’t care what other people say but then it is it has an effect even on the players they are not going to get the fans at Sansiero to make that noise about Inter today is going to be a different issue altogether

<$B$> well let’s uhm well with me I don’t think the crowd is going be a problem

<$A$> Inter wins but

<$D$> we shall see uh latest by lets say ha

<$A$> it’s my prayers that Chelsea will win by three goals to one

<$B$> well even if Chelsea is going to win it’s not going to be I mean with a high margin just a minute I mean difference

<$A$> whe* whether high margin or small margin it’s my prayer that Chelsea will win

<$D$> I’m staking my doubt from the two two goal interval will work for Chelsea

<$A$> I I I I I

<$B$> uhm I think I think it’s going to be a I think the game is uhm well <.>cheals</.> well uhm for Chelsea to have the best results it’s going to be on I mean on the penalty scores

<$C$> that is a prediction

<$D$> yeah what are you saying here

<$B$> yeah it’s a

<$A$> it’s going to be what

<$B$> on the penalty spot

<$A$> nah

<$D$> this match is not it’s it’s a one touch win for Chelsea it’s a one touch win for Chelsea

<$A$> this match will end in ninety minutes it will end within ninety minutes this match

In line 3 of the extract speaker <$D$> introduces BE GOING TO, which is taken up by the subsequent speaker <$B$> in the next turn. In line 15 speaker <$A$> reintroduces WILL, which is interrupted by speaker <$B$> with BE GOING TO in the following turn. Speaker <$A$>, however, sticks to WILL, which is continued by speaker <$D$>. Speaker <$B$> continues with BE GOING TO in his next
turn in line 23. This time, in reacting directly to speaker <$B>'s comment, speaker <$A> also sticks to BE GOING TO. A few lines later he uses WILL again. This example shows that while there is still considerable variation between WILL and BE GOING TO even within one part of the file, there is a tendency for individual speakers to stick to the FTE that they have used before. However, when reacting directly to another speakers’ utterance, speakers tend to use the FTE used by the previous speaker.

What is the reason for these tendencies? First of all, persistence or repetition is a phenomenon that has been noticed by many corpus-linguists when studying morphosyntactic variation (e.g. Poplack 1980; Poplack & Tagliamonte 1993; Weiner & Labov 1983; Szmrecsanyi 2006). Szmrecsanyi’s (2006: 109–129) multivariate analysis of persistence-related and non-persistence-related factors in the choice of WILL and BE GOING TO in several spoken native varieties of English shows that “speakers have a very marked tendency to avoid switching between future markers” (2006: 127). More precisely, he notes that recency of use is a significant factor, especially if the future marker was produced in the same turn and by the same speaker. The situation in CS-GH represented a mixed picture: While some speakers generally stick to one FTE, other speakers also show a high degree of convergence toward their interlocutors in conversations. Judging from the predictions of the regression models above, factors related to individual speakers and/or discourse events explain more of the variation of FTEs in CS-GH than in spoken ICE-GB.

6. Discussion

As was shown in this paper, in spoken GhE WILL and BE GOING TO generally favor the same linguistic environments as in spoken BrE, but in the former they are not as clearly distributed syntactically and semantically, with sentence type being the only significant constraint. In spoken BrE, BE GOING TO is highly grammaticalized and becomes more and more frequent in those contexts in which it is interchangeable with WILL (cf. also Mair 1997). WILL is more and more pushed back into 1st person, declarative contexts, most likely as the result of the I’ll-collocation. However, while most constraints seem to be of a lexical and syntactic nature (cf. also Torres Cacoullos & Walker 2009), WILL seems to persist in its original meaning of ‘willingness’ or ‘volition’ in highly specific contexts, reserving slots like I won’t and won’t you for interpretations of ‘refusal’ and ‘offer’.

In spoken GhE, the variant forms of WILL and BE GOING TO are not to be taken as heavily loaded forms in terms of syntactic, lexical, and pragmatic functions. These differences are most likely due to the nature of the (historical) input
and the mode of acquisition and use of English in Ghana. Language contact and the nativization of English in New English contexts may lead to the transformation of variable systems on the levels of syntax/lexis, semantics/pragmatics, and style or other extralinguistic dimensions. However, the present study has shown that major differences are not found on higher levels, such as the syntactic and semantic predictors chosen for analysis, but on lower levels. This is in line with more recent studies by Szmrecsanyi and colleagues on probabilistic grammars in World Englishes (cf. Szmrecsanyi 2015). Transfer of patterns from the substrate languages into English might eventually lead to the emergence of novel constraints, for example, the avoidance of BE GOING TO with motion verbs. As E. Schneider states,

> innovations and distinctive structural properties of PCEs [Postcolonial Englishes] are frequently positioned at the interface between lexis and grammar, i.e. certain words but not others of the same word class prefer certain grammatical rules or patterns.

(2007: 83)

It will have to be shown in future studies whether the strength of constraints in the future time system of New English varieties correlates in any way with the developmental stage in the sense of the Dynamic Model. The currently fewer constraints in the variable system of GhE than in BrE and the extreme differences between individual speakers in the choice of future time expressions mark GhE as a very heterogeneous variety and could also be seen an indicator for its status as an L2 variety.

In language contact situations, as much as in adult second language acquisition scenarios in general, patterns of variation in a system are not always fully acquired, but at times include only certain constraints. While the reasons for this may be manifold, one important aspect is certainly the nature of the target language (cf. Bayley 2005: 4–5). As Bayley suggests, “we must compare the pattern of variation in learner speech with the pattern of variation in the vernacular dialects with which learners are in contact and which they may select as the target” (2005: 5). In the case of the present data, we must assume that the patterns of variation apparent in the spoken, informal part of ICE-GB are not present in the kind of English that all Ghanaian speakers of English are primarily exposed to. Nevertheless, these patterns of spoken English are exactly those that shape the variable system of FTEs. English in Ghana is, on the one hand, acquired via formal education in schools and, on the other hand, via grassroots spread from other speakers of English within the same speech community. Input from native varieties of English is present, of course, but more often through various types of public media than through direct contact. The particular constructions that are important for the shape of the variable system in one variety of a language might not necessarily be important in another variety of the same language. Most of
the conventionalized expressions that are deeply entrenched in the native variety and that are responsible for the emergence of given constraints do not have the same effect in another variety. As BrE and GhE also differ drastically with respect to their speakers’ discourse conventions (e.g. politeness strategies, cf. Schneider 2015; Anderson 2013), we can assume that these differences account for differences in patterns of variation to some extent. In classroom situations, lexical, syntactic and semantic constraints are usually only explicitly taught if a failure to observe them would result in ungrammatical structures. While FTEs constitute a topic of grammar that is discussed in schools, discourse conventions and fixed expressions are typically transmitted in the immediately relevant context only. Many English teachers in secondary schools in Ghana are not familiar with idiomatic usage of BrE or AmE. Furthermore, a local variety it likely to stick to its own discourse conventions which are more entrenched the more frequently they occur. Less semantic and pragmatic load on individual variant forms of the FTEs WILL and BE GOING TO and the concurrent redistribution of predictors in the variable system in favor of cognitive or psycholinguistic constraints (syntactic/lexical, frequency- and/or persistence-related factors) might be seen as the result of such types of language contact scenarios.

What did not become clear from the data is whether the choice between WILL and BE GOING TO represents a stable situation or ongoing change. At this point it can only be concluded that the generally more frequent use of WILL in GhE points to a more conservative system, possibly shaped by written usage and the type of English taught in Ghanaian schools. Increased use of BE GOING TO by some speakers, by contrast, seems to be a local development rather than a direct copy from BrE or AmE. However, as Mair suggests, “with very few exceptions regional contrasts in the use of modals and semi-modals across varieties of English are embedded in more important diachronic drifts and are therefore generally temporary and ephemeral” and may often rather represent genre differences than clearly regional ones (2015: 141). Future research into the diachronic development of GhE may shed light on that matter.

Emerging constraints may also be of a sociolinguistic nature: Among GhE speakers, unlike be going to, be gonna is strongly associated with AmE and, if used by a GhE speaker, dubbed as LAFA (locally acquired foreign accent; Bruku 2010, Shoba et al. 2013). The (non-) use of be gonna in GhE could thus consciously or unconsciously be triggered by attitudinal factors, whereas in BrE and AmE its use is primarily determined by style. Huber (2014) reports differences between written GhE and BrE with respect to stylistic variation in the relativization system and concludes that in some cases the social prestige of linguistic variants adopted from the input variety is neutralized. While research on the acquisition of sociolinguistic variables has been carried out in the area of second-language acquisition
studies, it has thus far been largely neglected in New English studies and will have to be more thoroughly investigated in future studies (for potentially interesting variables, cf. Schneider 2015).

In studying the sociolinguistic meaning of variables it will increasingly become important to consider intravarietal differences within New Englishes. Unfortunately, the ICE-components are too small to compare subgroups of speakers, especially if researchers want to go beyond long-established categories such as age, gender and education. Specifically fruitful seem to be new sociolinguistic categories such as “extent of contact to native speakers” or “domains of usage”, i.e. text-type proficiency (for a discussion cf. Schneider 2015). Genre differences, as well as interspeaker differences based on new variables, will ultimately contribute to the understanding of the complex processes at work in the emergence of a New English variety.

References


### Appendix

#### Table A.1 Mixed-effects linear regression models fitted to the data

<table>
<thead>
<tr>
<th>Future Time Expression</th>
<th>Spoken ICE-GB</th>
<th>CS-GH</th>
</tr>
</thead>
<tbody>
<tr>
<td>WILL</td>
<td>266.66 (480) = 62.90%</td>
<td>358.15 (505) = 76.51%</td>
</tr>
<tr>
<td><em>will</em></td>
<td>53.88 (97)</td>
<td>222.69 (314)</td>
</tr>
<tr>
<td><em>'ll</em></td>
<td>189.44 (341)</td>
<td>112.05 (158)</td>
</tr>
<tr>
<td><em>won’t</em></td>
<td>23.33 (42)</td>
<td>23.4 (33)</td>
</tr>
<tr>
<td>WILL + Progressive</td>
<td>4.44 (8) = 1.04%</td>
<td>22.69 (32) = 4.84%</td>
</tr>
<tr>
<td><em>will</em> + Progressive</td>
<td>1.11 (2)</td>
<td>17.02 (24)</td>
</tr>
<tr>
<td><em>'ll</em> + Progressive</td>
<td>3.33 (6)</td>
<td>5.67 (8)</td>
</tr>
<tr>
<td>BE GOING TO</td>
<td>152.77 (275) = 36.04%</td>
<td>87.23 (123) = 18.63%</td>
</tr>
<tr>
<td><em>be going to</em></td>
<td>150 (270)</td>
<td>83.68 (118)</td>
</tr>
<tr>
<td><em>be gonna</em></td>
<td>2.77 (5)</td>
<td>3.54 (5)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>423.88 (763)</td>
<td>458.33 (660)</td>
</tr>
</tbody>
</table>
CHAPTER 6

Ongoing changes in English modals
On the developments in ELF

Mikko Laitinen
University of Eastern Finland

The article investigates how ongoing grammatical change, widely documented in various native varieties, is adopted in advanced lingua franca use of English (ELF). It incorporates a broader perspective on ELF than previously, seeing it as one stage in the long diachronic continuum of Englishes rather than as an entity emerging in spoken interaction. The first part details a corpus project that produces written multi-genre corpora suitable for real-time studies of how ongoing variability is reflected in lingua franca use. It is followed with a case study investigating quantitative patterns in a set of core and emergent modal auxiliaries. The results suggest that in cases of substantial recent changes in the core varieties of English, lingua franca uses polarize the diffusion of change. The conclusions suggest that a diachronically-informed angle to lingua franca use offers a new vantage point not only to ELF but also to ongoing grammatical variability.

Keywords: ongoing change, modal auxiliaries, emergent modal auxiliaries, English as a lingua franca, second language use

1. Introduction

This article discusses new ways of investigating the ongoing expansion and diversification of English in the expanding circle.1 It focuses on advanced non-native use, i.e. situations in which English is adopted as an additional resource alongside people’s first languages (L1) in countries which do not have colonial links to Britain (Schneider 2014). The term English as a lingua franca (ELF) is used as a label for such uses. The article looks into the role of diachronic processes of change

---

1. I wish to thank the reviewers for their valuable comments on an earlier version of this article. The usual disclaimers apply.
and compares to what degree the patterns of variability observed in ELF differ from those in native English varieties.

The approach is motivated by (a) the recent calls in the ELF paradigm to provide diachronically-informed evidence of placing it in a diachronic context (Seidlhofer 2011: xi), and (b) the research on bridging the paradigm gap between second and foreign language use (Mukherjee & Hundt (eds) 2011). The latter has suggested that the increased globalization requires innovative empirical studies of how the established native varieties and the new Englishes overlap, and scholars have called for material development and larger non-native corpora (Hundt & Vogel 2011).

The case study in Section 4 investigates how ongoing grammatical changes, which are widely documented in various native English varieties, are adopted in ELF. This approach incorporates a broader perspective on ELF than previous studies, which have primarily focused on meaning making in spoken interaction (cf. Jenkins, Cogo & Dewey 2011; Seidlhofer 2011; Mauranen 2012). Rather than considering ELF as an ideological entity emerging from interaction, it is seen as one stage in the long diachronic continuum of Englishes. A similar focus has recently been suggested in the study of indigenized L2 Englishes as Nöel, van Rooy and van der Auwera (2014) have called for the study of world English grammar from a diachronic perspective.

The changes discussed concern the recent and ongoing changes in the modal and emergent modal auxiliaries, documented in the history of English (Krug 2000) and in the recent past in various native and second language varieties (Leech 2003, 2011, 2013; Collins 2009; Millar 2009; Nöel, van Rooy & van der Auwera 2014). On the methodological level, the study not only repeats what has been done but develops the empirical basis of corpus-based research of English in the expanding circle. This means collecting new corpus sources and broadening the scope of discourse situations and genres included. The existing non-native corpora (viz. both the learner corpora and the primarily spoken ELF evidence) contain a narrow selection of genres, making them not ideal material for diachronic analyses. As is well known, the study of language change builds on the idea that textual stratification is vital for understanding the two pathways of change, i.e. change from below and change from above (cf. Rissanen 2008; D’Arcy & Tagliamonte 2015).

ELF is not a focused language variety, yet the role of diachronic processes shaping it is both theoretically and empirically relevant (Laitinen forthcoming). A case in point is the ongoing grammatical change and its underlying sociolinguistic

---

2. Similarly, research on learner English has focused on interlanguage phenomena and proficiency in acquisition and not on time and diachronic processes (Granger 2008; Granger, Gilquin & Meunier (eds) 2015).
processes. They include colloquialization, the gradual shift of spoken norms to
from traditional learner corpora, points out that learner varieties exhibit “regis-
ter interference” in the forms of overrepresentation of speech-like characteristics.
Since many genres in which English is used in non-native contexts are informal
mediated online texts, it should mean that ongoing grammatical changes are ac-
celerated by speakers/writers in today’s globalized linguistic marketplaces. In ad-
dition, this research contributes to the study of ELF which has so far focused on
a limited set of grammatical features (e.g. dropping the 3rd person -s, omitting
articles, and leveling in the relativizers). Much less attention has been placed on
broad quantitative drifts in grammar.

The structure is such that Section 2 presents the material and methodological
considerations. Section 3 is an overview of the changes in the modal paradigm,
and Section 4 presents the results.

2. ELF on a diachronic continuum

This section first discusses some theoretical and methodological considerations
and illustrates what kinds of insight can be gained by investigating ELF on a dia-
chronic continuum. Second, it presents the corpora used in the case study.

2.1 Some theoretical and methodological considerations

One result of ongoing globalization is the fact that a substantial bulk of use of
English today is non-native communication of varying levels of proficiency. Non-
native speakers outnumber native speakers, so it is justified not only to ask what
happens in this process but also what types of similarities and differences exist
between ELF and the native varieties. The role of multilingual context in shaping
English also calls for more research; Mair (2013) has suggested that the farther a
language travels from its roots, the more likely it is affected by the multilingual
settings in which it is used.

Research in (post)-colonial contexts has shown that the spread of English
is characterized by extraterritorial conservatism (Trudgill 2004), according to
which contact means adult language learning, which in turn involves simplifica-
tion. Recently, Hundt (2009) has suggested a complex typology of grammatical
change in colonial settings, involving not only true colonial lag, but also various
scenarios which emerge from empirical studies (such as extraterritorial innova-
tion, truly divergent patterns, and kick-down developments, which include the
changes in the modals for instance). In the concentric circle model, the expanding
circle contexts are seen as norm dependent, relying on norms of usage and pro-
scriptions from the inner circle (Mesthrie & Bhatt 2008: 29).

In addition to the sociolinguistic processes in ongoing change (colloquializa-
tion for instance), the present study is informed by social network theory, which
predicts that weak social ties promote diffusion of innovations (Milroy & Milroy
1985). The working hypothesis is that multilingual individuals who adopt English
as an additional resource alongside their L1s and engage in EFL communication
have on average more weak ties than the rest of the population, and they therefore
act as agents of change in diffusion. The influence from multilingual settings could
therefore lead to accelerated change at least in certain non-native communities.
Compare this to the Civil War effect proposed by Raumolin-Brunberg (1998: 367–
368), according to which mobility and subsequent weak ties during the English
Civil War contributed positively to grammatical change in England.

The methods are drawn from short-term diachronic linguistics in English
(Leech et al. 2009: 24–50). The field, according to Mair (2009: 1109), relies on
“the same principles as all historical corpus linguistics” and the core is to under-
stand how language changes takes place. The unit of analysis typically consists of
variables, two or more ways of expressing the same meaning (Rissanen 2008) or
in some cases the studies focus on fluctuations in frequencies on the token level
which do not have to have variant forms (e.g. the progressive). Whatever the fo-
cus, the forms are investigated with reference to both internal processes of change,
such as grammaticalization, metaphorical extension and semantic change, and to
various external factors, which may include speaker-related, regional or genre-
based independent variables.

2.2 Material

At the level of evidence, the study of recent history of English draws from system-
atically-collected corpora that offer equidistant observation points at certain inter-
vals (Leech 2013). The various Brown siblings offer equivalent observation points
from 30-year intervals from the early 20th century to early 2000s and contain
texts from a range of genres, thus providing tools for tracing diachronic variability.

The situation is different on the non-native side. Space does not allow a com-
prehensive analysis of the various learner and ELF corpora, but suffice it to say that
the existing corpora have by and large been designed for synchronic studies. New
multi-genre corpus resources are needed, and the author is currently compiling
corpus resources on the following broad principles:

3. The BE06/AE06 corpora are exceptions, offering an observation interval of circa 15 years.
a. the corpus resources should primarily cover written genres as the majority of
   the existing ELF corpora contain spoken material;

b. they should take into account the diversity of English uses and be based on
   bottom-up style corpus design, which sets out by looking at what types of
   genre exist;

c. they should enable quantitative comparisons with the existing corpora from
   which much of the evidence of the recent history of English is drawn;

d. the corpus design should enable diachronic and diatopic approaches to ex-
   ploring how recent and ongoing changes are adapted in ELF use.

The corpora go with their working titles of the Corpus of English in Sweden (SWE-
CE) and the Corpus of English in Finland (FIN-CE). As opposed to learner corpora,
they cover English in use in non-instructional environments. The two countries
are used as a case study because the role and nature of English there is currently
undergoing changes, which are well documented (Leppänen et al. 2011; Bolton
& Meierkord 2013). In addition, the main languages of these two countries, i.e.
Swedish and Finnish, are typologically different.

The corpus design is based on the idea that the corpora will be small and tidy
(Hundt & Leech 2012), and the target is one million words per corpus. During the
compilation, all the informants are identified, and the compilers have knowledge
of the extent to which the materials have been subjected to normative language
checking by professional editors/translators and native speakers. Preference is
given to texts that are not edited, but it is assumed that the more informationally
oriented a text is, the more likely it is to have undergone some degree of language
checking and collaborative effort. Materials edited by native speakers are excluded.

In the case study, the written material consists of 1,050,649 words of texts,
drawn from three genres (see Table 1). In the formal end, there are (a) 456,510
running words of information-dense academic prose texts. They consist of single-
authored student theses from disciplines other than language studies, and were
collected from publicly-available databases. The collection consists of samples
of 3,000–5,000 words from the introduction and conclusion sections. The news

---

4. No large scale language shift is taking place in these two countries, but the phenomenon
could best be described as urban multilingualism involving English. For instance, Leppänen et
al.'s (2011) large-scale survey results establish three macro-level groups in Finland: Some 16% of
Finns could be described as “have-it-alls” when it comes to attitudes to and uses of English. In
this group, English has strong presence as a result of extensive studies, self-reported proficiency
and extensive use. The group consists of younger respondents (85% aged under 45), who live in
urban areas, have a university or polytechnic degree and who work as managers or experts (the
two other groups being “haves” and “have-nots”). Unfortunately, no similarly extensive survey
data are available from Sweden yet.
sample consists of (b) 164,271 words of news texts written by journalists for online news platforms. As opposed to the uptight academic genres, news texts in general are described as agile and economic by Hundt & Mair (1999). Non-fiction texts are represented by (c) 429,868 words of personal/professional blogs, freely available in the blogosphere. According to Sand (2013), blog texts in general represent vernacular, spoken-like written texts.

The spoken results are drawn from the parts-of-speech tagged Vienna-Oxford International Corpus of English (VOICE), containing 1,023,082 orthographically defined words from roughly 110 hours of recording. The recordings are naturally-occurring and non-scripted face-to-face interactions in which ELF is used, and the data were collected between July 2001 and November 2007 <http://www.univie.ac.at/voice/page/corpus_information>. The corpus is available on-line.

These ELF corpora are synchronic, but similarly to Collins (2009) it is assumed that the patterns of variability in the channel (spoken–written) and genre, when compared with other corpora, can be used to make diachronic conclusions. Such an approach requires that the corpora (their corpus design and sampling frames) should ideally be as equivalent as possible to the corpora used as the point of comparison but take into account the considerable differences in (1) the sociolinguistic situations and (2) the output/production of texts in non-native contexts (Laitinen 2016; Laitinen & Levin 2016).

3. Notes on the ongoing changes in the modal system

The modals in Section 4 are based on Leech (2013), and the selection criteria are discussed extensively in Leech et al. (2009: 92–98). The relevant instances of core modals have the four N(egation)I(nversion)C(ode)E(mphasis) properties, and include can, could, may, might, will, shall, must, should, would, ought (to), and need (n’t) followed with a bare infinitive (Leech et al. 2009: 94). The emergent modals consist of a heterogeneous group of modal idioms, semi-auxiliaries and main+non-finite verb structures. According to previous studies (Collins 2009; Leech 2013), they have four main properties: They have emerged through grammaticalization in which full lexical items have gained grammatical properties.
They have idiomatic meanings, which have emerged through metaphorical extension, as shown by Krug (2000: 55) through: *I have a letter to write* > *I have to write (a letter)*. In addition, they are semantically related to core modals, expressing meanings related to non-factuality, such as possibility, necessity, obligation, ability, and permission. Lastly, their emergence can be attributed to colloquialization, and their increase in native spoken genres is considerably steeper than in written ones (Leech 2013). The items included are be going to, have to, (have) got to, want to, need to, be able to, (had) better, and be supposed to, and they also include the contracted forms of gotta, gonna, hafta, and wanna.5

Two recent quantitative trends emerge in previous studies. On the one hand, the core modals have declined substantially in the second half of the 20th century (Leech 2003, 2013). The decline, traceable to the early part of the century, continues in the 21st in the web-based 2006 versions of the Brown family of corpora. It is led in American English (AmE) which is “roughly one generation in advance of the British decline” (Leech 2013: 100). At the individual modal level, the AmE data show consistent decline in “every single modal” whereas in British English (BrE) two items, can and could, increase slightly in frequency (Leech 2013: 100).

On the other hand, emergent modals have increased significantly. This increase is however not as fast as the decrease in the core modals, and Leech (2013) suggests “modality deficit” in the present-day written English genres. These quantitative trends are also supported by the results from the Corpus of Historical American English (Leech 2013), and the evidence from spoken language suggests that these developments are changes from below as both the decline and the increase are steeper in spoken than in written texts.

These quantitative developments are nevertheless far from straightforward. Bowie, Wallis and Aarts (2013: 90), investigating recent changes in spoken genres, show that not all core modals are declining at the same rate, and some modal items are undergoing slight increases. In addition, in written evidence, the most frequent core modals (will, would, can) have been least subject to the drop; the mid-frequency (e.g. could, may, should) have decreased slightly, whereas the least frequent ones (shall, need, ought) have been subject to substantial decreases. Similarly, in the emergent modal side, the largest increases have taken place with a small group of items. For instance, need to, want to and be supposed to have undergone substantial increases, whereas (had) better and (have) got to have decreased in frequency (need to in contemporary BrE is extensively discussed in Nokkonen 2015).

---

5. The small caps indicate that the items are lemmas and the queries included all the possible forms for the lemmas.
In addition to the core native varieties, several other varieties have also been investigated. Collins (2009) compares nine varieties, consisting of both native and non-native inner and outer circle Englishes, and concludes that as regards the developments in both modal categories, the diachronic developments are led by AmE. His results highlight the role of spoken AmE in which the rise of the emergent modals is substantial; in addition, he shows that the South-East Asian varieties lead the way in the outer circle. Collins’ (2013, 2014) diachronic studies show that Australian English is evolving in the same direction as the inner core varieties, but he also shows that speakers tend to disprefer modals even more but not adopt the quasi-modals as eagerly, which he explains is brought about by endonormative independence. As for individual modal structures, Filppula’s (2012) findings of will/shall be V-ing suggest that BrE, surprisingly, seems to be following the various outer circle varieties.

4. Results

The results are presented in three sections. 4.1 illustrates the broad patterns and focuses on genre differences, which, in the absence of multi-genre lingua franca evidence so far, have received limited attention. 4.2 focuses on the frequency-based order of individual modal and emergent modal types in the data. The ELF evidence is set side-by-side with Leech et al.’s (2009: 283, 286) findings from standard edited written English in the FLOB (BrE) and Frown (AmE) corpora and with Leech’s (2013: 112) results from the spoken demographic subcorpus of the British National Corpus (BNC) and the Longman Corpus of Spoken American English (LCSAE). 4.3 presents evidence placing lingua franca use on the diachronic scales of inner and outer circle varieties, which have been proposed in Collins (2009).

The results are normalized per 1 million words, and they focus on quantitative, form-based patterns. These results give room for qualitatively-oriented analyses of micro-level patterns of epistemic, deontic and dynamic meanings, which are left for future studies.

---

6. FLOB and Frown are roughly 1 million words each; the demographic part of the BNC is 4.2m words, and LCSAE is 5.1m. The AmE and BrE results exclude the figures for be to. Unfortunately, Leech (2013) does not include the absolute figures for BE06/AE06, and hence the comparisons are based on quantitative evidence from the 1990s versions, FLOB/Frown. If one estimates Leech’s (2013: Figure 1) results, the ratios for the 2006 versions are only slightly lower than the 1990s ratios and do not change the broad pattern in Table 2 here.
4.1 Frequencies and comparisons with the native varieties

Table 2 shows the normalized frequencies of modals and emergent modals in lingua franca use, drawn from the written FIN/SWE-CE and the spoken VOICE material. The total frequency of the core modals is 30,247 instances, and that of the emergent modals is 9,880. It results in the ratio (core/emergent) of 3.1, meaning that for each incoming emergent modal, the frequency of core item is roughly three times higher. When compared with Leech’s (2013) results in the two main metropolitan varieties, the ratio lands in between the more advanced AmE (with the ratio of 2.5) and BrE (3.6) varieties. A contingency table test shows that the differences between ELF and both of the native varieties are statistically highly significant ($p < .001$).

<table>
<thead>
<tr>
<th>Variety (corpora used)</th>
<th>Ratio</th>
<th>Core modals</th>
<th>Emerg. modals</th>
</tr>
</thead>
<tbody>
<tr>
<td>AmE (LCSAE + FROWN)</td>
<td>2.5</td>
<td>32,188</td>
<td>13,047</td>
</tr>
<tr>
<td>BrE (BNCdemog + FLOB)</td>
<td>3.6</td>
<td>34,621</td>
<td>9,610</td>
</tr>
<tr>
<td>ELF (VOICE + FIN/SWE-CE)</td>
<td>3.1</td>
<td>30,247</td>
<td>9,880</td>
</tr>
</tbody>
</table>

As expected, ELF use follows the norm-providing leader (AmE) in this particular change. What is noteworthy however is that the non-native ratio is more advanced than BrE. The difference seems to be brought about by the substantially lower frequency of core modals in the ELF material, much more so than the frequency of emergent modals.

This first result should be interpreted with some degree of caution since there are qualitative differences between the sources of evidence, and I am not comparing exactly like-with-like here. In more particular, the written materials of the native varieties come from edited and published sources solely, and are not directly comparable with the ELF evidence, which are primarily web-based (but see the discussion of such materials below in this section). Yet, it should be pointed out that the written lingua franca corpus contains evidence from various genres, both formal and informal, as shown in Section 2. There is also some degree of diachronic difference as the native evidence comes from the 1990s whereas the ELF results are from the 2000s and 2010s. I will return to the questions of diachronic depth and the role genre later, but it is fair to say that the result merits further study, since it goes against Collins’ (2009) findings in the International Corpus of English (ICE). The non-native varieties which he investigates have substantially lower proportions of emergent modals than any of the native varieties used in the study.

The next result concentrates on the impact of genre in the ELF corpora. Previous studies suggest that both the decline in the core modals and the increase
in the emergent ones are steeper in the spoken mode (Leech 2013: 105–107; Bowie, Wallis & Aarts 2013). There are two angles to this issue. Firstly, there is the need to shed light on the extent to which the new media texts, with a greater likelihood of spoken-like characteristics, skew the overall results above. Our informal observation while collecting the material is that the blogosphere is increasingly characterized by high-stakes writing in the sense that various individuals write professional blogs. They might receive income from advertisements based on the number of visitors to the pages, which could mean that more attention is paid to text production. Secondly, despite the fact that blogs are not represented as a text type in the native corpora used as points of comparison, they are a text type which cannot be ignored when studying non-native English use. As pointed out in Section 2, English is often used as the medium in electronically-mediated texts, which themselves are an interesting topic for variationist study.7

Figure 1 illustrates the patterns of situational variation in the three written genres and in the spoken VOICE results. It shows the ratios of core modals divided by emergent modals (the columns) and also includes the proportional share of the emergent modals in each text type (the line graph on the secondary y-axis on the right). No comparisons with native materials are made, since fiction, which is the most “modal rich” text type in Leech et al. (2009: 75), does not really exist as a text type in ELF uses according to our knowledge.

Figure 1. Patterns of situational variation in written and spoken ELF corpora

7. One objective in this research is to produce comparable ELF materials so that in 15–20 years’ time we could make reasonable statistical comparisons between materials from 2010s and 2030s, investigating change in ELF in real time.
In the written corpus, the frequencies show a clear pattern in which the core modals are most frequent in the informationally-oriented educational texts, and the frequency decreases in news and in new media texts. The order of the emergent modals is the opposite, and they are nearly 40% more frequent in news than in education, and over twice as frequent in new media texts. These differences are visible in the ratios (the primary y-axis) and the proportional frequencies (the secondary y-axis). For every incoming emergent modal in education, there are nearly six core modals. In news, the ratio is roughly 4:1, and in new media texts, the share of emergent modals actually surpasses that of spoken VOICE so that the ratio is lower and subsequently the proportion of the emergent modals is larger in new media texts than in the spoken corpus here. The log-likelihood values indicate statistically highly significant differences between the two written genres (education, news) and the new media and spoken VOICE ($p < 0.0001$).

These results are new not only in terms of the quantitative patterns but also methodologically. So far, much of the previous research on electronically-mediated genres and discourses has been synchronic, but these results suggest that a diachronic perspective yields new information on the role of mediated genres in contributing to ongoing language change. Similarly, in the previous corpus-based studies on the globalization of English, the impact of mediated genres is an understudied phenomenon. One question to be explored in more detail in the future is whether the emergence of new media texts accelerates ongoing change, as the results here suggest. If one zooms in on the evidence in the new media subcorpus alone, the findings indicate that among the ten most frequent items, there are three emergent modal types (want to (wanna), have to (hafta) and need to). The most frequent emergent modals, have to (hafta) and want to (wanna) are indeed more frequent than all the other core modals except the four most frequent ones (can, will, would and could). The next section therefore concentrates on the individual modals and their frequencies.

4.2 Quantitative patterns in individual core and emergent modals

As pointed out in Section 3, the decreases of the core modals and the increases in the emergent modals are not uniform but contain substantial variation depending on the type. Figures 2 and 3 visualize the orders of frequency of modals in the written corpora and compare them with the evidence from AmE and BrE, as reported in Leech et al. (2009: 283).

Figure 2 shows that the three high frequency core modals are high frequency items also in ELF. A noticeable pattern is that the order is reversed as can is the most frequent one in non-native use and is followed by will and would. The log-likelihood values show that the observed differences are statistically significant.
(at 0.01% level) with can (the log-likelihood value 404.31, p < 0.01%) and would (333.55), but not with will (12.40). The differences call for an explanation and a second-language-acquisition-oriented way to understand them would be to suggest that non-native speakers/writers overuse speech-like grammatical features, such as core modals. Yet, not all core modals are overused, which calls for alternative explanations.

The explanation proposed draws from the idea of textual variation and makes a connection between the synchronic quantitative patterns observed in lingua franca use and the general diachronic shifts in the established native varieties. Out of the 11 core modals, can/could are the only ones showing increases in BrE and only slight decreases in AmE (Leech et al. 2009: 283). Can in BrE exhibits a slight increase in the second half of the 20th century, and this increase is reflected in proportionally larger share in ELF use.

Figure 2 shows further evidence of the polarizing effect of ELF in relation to the recent history of English. Of the mid- and low-frequency core modals, the shares of those items that have undergone the largest decreases in the native varieties seem to be even lower in ELF. Among them there are, for instance, may and must whose frequencies are statistically highly significantly lower (LL values 71.39 and 148.52 respectively). It is noteworthy that the decrease of must, primarily in deontic uses, in the core native varieties seems to be more pronounced in ELF, as it is falling out of use. To what extent this is reflected in the proportions of epistemic and deontic uses is a topic to be tackled in future studies. What is however noteworthy, is that its frequency is not far from the low frequency modals (shall, ought to, need (n’t)), and when compared with the frequencies of emerging modals
(discussed below), it is less frequent than the majority of these items. The low frequency modals have nearly disappeared from non-native use.

Two additional core modals are worth mentioning since they are undergoing less dramatic decreases in the inner core varieties than the “most notable losers” (Leech et al. 2009: 73), yet their decreases are more pronounced than those of the strongest items. As shown in Figure 2, should and might are more frequent in ELF than in native use. The differences are statistically significant for might (LL 15.97) but not for should.

Figure 2 presents a set of results for which polarizing effects were suggested to play a role. That is, for those core modals whose decreases were mild (or nonexistent as in the case of can), their quantitative patterns were accentuated in ELF, and the opposite, according to which the decreasing forms were further decreased, also held. I will next move on to presenting the quantitative patterns for emergent modals, and if this working hypothesis holds, we should see increases nearly across the paradigm, the exceptions being (have) got to and (had) better, both of which are decreasing in the core native varieties.

Figure 3. Order of frequency of emergent modals in BrE/AmE and in ELF

Figure 3 shows the normalized frequencies of the emergent modals in the same dataset. It reflects the fact that these items are in general more frequent in the ELF corpora than in the edited written FLOB and Frown. What is noteworthy, however, is the distribution of the frequencies of the individual items as they are greater in all the items but in (have) got to and (had) better. Of these, (had) better only appears six times in our 1-million word corpus, and all of the cases come from the new media subcorpus. They appear only with 1st and 2nd person subjects, (1).
(1) Sorry lads, this is a solid defense, you want to get into my team you **better start** performing!  
(FIN-CE, new media, 2009)

(HAVE) *got to* is slightly more frequent, but its frequencies are as low as the low-frequency core modals discussed above. The differences between non-native material and the native results in Leech et al. (2009: 286) are not statistically significant (LL 0.20).

The largest proportional differences are in *need to* (183%) and *want to* (86%), but *be able to* (83%) and *be supposed to* (82%) also show substantial differences. It is noteworthy that when these four emergent modals are seen on a recent diachronic continuum, all except one undergo considerable increases in the native varieties. Leech et al.’s (2009: 286) results show that with the exception of *be able to* (3% increase), all the others increase substantially in the late 20th century. The illustrations below show a selection of the relevant instances from the FIN/SWE-CE corpus.

(2) ‘I *need to* reclaim Swedishness’  
(SWE-CE, news, 2013)

(3) We need simpler rules and Member States *need not only to* implement the rules but also have them accepted.  
(SWE-CE, news, 2013)

(4) The hotel market is dominated by Scandinavian-owned chains, but we *want to* add more international chains.  
(SWE-CE, news, 2013)

(5) Unless you’re a billionaire who can afford to finance your own movie and do whatever you like, you need adjust. You don’t *want to* shock people so much that they leave the theatre and advise their friends against watching the film.  
(SWE-CE, new media, 2012)

(6) Even if you *are able to* search for anything and find an abundance of correct information on the internet you may simply choose to ignore what is not according to your view.  
(FIN-CE, new media, 2008)

(7) In the weekend I was accepted as iPhone Developer (and had to paid for that) and now I *am able to* develop nice programs  
(FIN-CE, new media, 2008)

(2)–(7) also illustrate a set of language-internal factors that will be subject to qualitative analyses in the future. *need to* is often used to express inherent necessity, often with an animate first person subjects, as in (2), or structures in which the origin of the necessity is obscured through the use of inanimate subjects (3). As pointed out by Krug (2000: 148–150), the diachronic path of *want to* has resulted in layering, in which the intentional readings (4) could be interpreted as less face-threatening alternatives to *ought to/must/have to*. In addition, the deontic senses
of obligation/necessity, as in (5) are also found in the material. In *be able to*, the majority of the uses point to personal ability (6), but there are also occasional cases of deontic meanings (7).

The contracted forms, such as *wanna*, do not explain the large differences between the native varieties and ELF. They only appear in the new media subcorpus, but with low overall frequencies, such as *(HAVE) got to* where half of relevant cases are contracted, as in (8).

(8) I cancelled dance practice today because of my mouth but I still *gotta go* over there to try out my first dress (!!) (SWE-CE, new media, 2012)

The results support the working hypothesis according to which considerable recent frequency changes in grammar (either increases or decreases) could actually be polarized in ELF. It seems that multilingual settings in today’s globalized world not only contribute to ongoing variability but also seem to accelerate diffusion. The last section investigates how ELF results complement the patterns of variability previously attested using a range of inner and outer circle evidence.

### 4.3 Diachronic fortunes of modals in ELF and in World English varieties

This section compares the observations in ELF with the results presented in Collins (2009). His study investigates to what extent the diachronic developments recorded in modals in the core varieties are “also occurring in the other Englishes of the world” (2009: 285). His material consists of eight ICE corpora and one specifically-tailored AmE corpus, which is a combination of select written texts in Frown and the spoken *Santa Barbara Corpus*. The primary division draws from Kachru’s (1985) concentric circles and is made between the inner circle (AmE, BrE, Australian and New Zealand English) and the outer circle Englishes, which he divides into Southeast Asian varieties (PhilippineE, SingaporeE and Hong Kong E), South Asia (IndianE) and one African variety (KenyanE). These varieties represent a continuum of Englishes so that while English is an official language in all the contexts examined, it is not always the case that it would be the L1 of the population.

Collins (2009) focuses on a set of core (*must, should, need (n’t), will and shall*) and emergent modals (*have to, (HAVE) got to, need to, be going to and want to*). The results show two tendencies: Firstly, the inner core variants, and AmE in particular, are leading the increase of the emergent modals and the decrease of the core items. Secondly, the development is most pronounced in the spoken medium, as the ratios of emergent modals in speech are substantially higher in the inner core varieties than in the various outer circle varieties.
His results are compared with the normalized frequencies (per 1m words) and the ratios in speech and writing in ELF. To make the corpus samples comparable, the spoken data come from the VOICE corpus and the written evidence from FIN/SWE-CE from which the new media texts are discarded. The sample size is 1,645,605 running words. Again, the sets of corpora are not based on identical corpus designs, yet, they are the closest one can get to one-to-one matching using the corpus resources available. For simplified visual presentation, Tables 3–5 show the total numbers from Collins (2009) but exclude the frequencies of individual items from the inner and outer circle varieties since they are available in his study. All the ELF results are included.

Table 3 below on p. 191 presents the figures of the core modals in ELF, and sets them side-by-side with the broad frequencies reported in Collins (2009: 285–286). If one accepts the attested decline in the core modals as the backdrop, these results add to the general conclusions in Collins (2009). The observations in ELF (6,188 tokens per 1 million words) fall within the range of the inner core (IC) results, but are not as advanced as them (5,815 tokens pmw). They are, however, well ahead of the Southeast Asian (SEA) and the other two Outer Circle (OC) varieties. As for the individual modals, the frequencies are particularly low for must, need and shall, all of which have undergone substantial decreases in the inner core varieties. The only modal category in which the frequencies are substantially higher is should, in which the frequencies are within the range of the SEA and OC varieties. Recall from Section 4.2 above that should has undergone less dramatic decreases in the native varieties than the other core modals, which is also reflected in its frequency.

To look at the frequencies in emergent modals, Table 4 presents the items investigated in the same format. The overall picture is similar, but the aggregate frequency is highest in ELF (4,518 token pmw). If one looks at the individual varieties, it is AmE and the modal-rich HKE in which there are higher total frequencies than in lingua franca use. As for the individual emergent modals, there are two items (viz. have to and need to) in which ELF has the highest frequency. For want to, the prominent role of HKE skews the picture.

These results partly confirm the previous results because they clearly show that both the developments are led in AmE. The core modal frequency is the lowest and that of emergent modals is the highest in that variety, and the inclusion of ELF data does not change this situation. However, they add a new dimension to the inner and outer circle results, suggesting that the empirical evidence from ELF complicates the canonized division of varieties.

To shed more light on the role of corpus data in explaining the evidence in this section, my last set of results looks into the ratios of the incoming emergent modals in spoken and written evidence. The working hypothesis is that if the ratio in the ELF evidence is substantially different from the results derived from the
Table 3. Frequencies of a set of core modals in ELF compared with the results in Collins (2009: 286)

<table>
<thead>
<tr>
<th></th>
<th>ELF</th>
<th>AmE</th>
<th>AUS</th>
<th>NZ</th>
<th>BrE</th>
<th>IC total</th>
<th>PHI</th>
<th>SIN</th>
<th>HK</th>
<th>SEA total</th>
<th>IND</th>
<th>KEN</th>
<th>OC total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Must</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>601</td>
<td></td>
<td></td>
<td></td>
<td>869</td>
<td>871</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Should</td>
<td>1,721</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,173</td>
<td>1,650</td>
<td></td>
<td></td>
<td>5,917</td>
<td>5,815</td>
<td>7,001</td>
<td>10,392</td>
</tr>
<tr>
<td>Need (n’t)</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td>32</td>
<td></td>
<td></td>
<td>32</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Will</td>
<td>3,951</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3,888</td>
<td>5,989</td>
<td></td>
<td></td>
<td>5,209</td>
<td>5,917</td>
<td>7,001</td>
<td>10,392</td>
</tr>
<tr>
<td>Shall</td>
<td>114</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>131</td>
<td>182</td>
<td></td>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6,188</td>
<td>5,319</td>
<td>5,741</td>
<td>6,284</td>
<td>5,917</td>
<td>5,815</td>
<td>7,001</td>
<td>8,774</td>
<td>10,392</td>
<td>8,722</td>
<td>7,525</td>
<td>6,190</td>
<td>7,976</td>
</tr>
</tbody>
</table>

Table 4. Frequencies of a set of emergent modals in ELF compared with the results in Collins (2009: 285)

<table>
<thead>
<tr>
<th></th>
<th>ELF</th>
<th>AmE</th>
<th>AUS</th>
<th>NZ</th>
<th>BrE</th>
<th>IC total</th>
<th>PHI</th>
<th>SIN</th>
<th>HK</th>
<th>SEA total</th>
<th>IND</th>
<th>KEN</th>
<th>OC total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAVE to</td>
<td>1,895</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,281</td>
<td></td>
<td></td>
<td></td>
<td>1,683</td>
<td>1,541</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(HAVE) got to</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>268</td>
<td>146</td>
<td></td>
<td></td>
<td>105</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEED to</td>
<td>605</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>359</td>
<td>450</td>
<td></td>
<td></td>
<td>382</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BE going to</td>
<td>735</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,437</td>
<td>918</td>
<td></td>
<td></td>
<td>774</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WANT to</td>
<td>1,251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,031</td>
<td>1,276</td>
<td></td>
<td></td>
<td>1,081</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4,518</td>
<td>5,869</td>
<td>4,216</td>
<td>3,636</td>
<td>3,777</td>
<td>4,375</td>
<td>3,692</td>
<td>4,331</td>
<td>5,398</td>
<td>4,474</td>
<td>3,059</td>
<td>2,934</td>
<td>3,883</td>
</tr>
</tbody>
</table>
Table 5. Ratios of emergent modals in speech and writing (the inner and outer circle ratios from Collins (2009: 286))

<table>
<thead>
<tr>
<th></th>
<th>ELF</th>
<th>AmE</th>
<th>AUS</th>
<th>NZ</th>
<th>BrE</th>
<th>IC total</th>
<th>PHI</th>
<th>SIN</th>
<th>HK</th>
<th>SEA total</th>
<th>IND</th>
<th>KEN</th>
<th>OC total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAVE to</td>
<td>3.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.98</td>
<td></td>
<td></td>
<td></td>
<td>2.26</td>
<td></td>
<td></td>
<td>2.49</td>
</tr>
<tr>
<td>(HAVE) got to</td>
<td>31.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.18</td>
<td></td>
<td></td>
<td></td>
<td>4.59</td>
<td></td>
<td></td>
<td>4.10</td>
</tr>
<tr>
<td>NEED to</td>
<td>1.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.24</td>
<td></td>
<td></td>
<td></td>
<td>0.89</td>
<td></td>
<td></td>
<td>1.13</td>
</tr>
<tr>
<td>BE going to</td>
<td>9.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.14</td>
<td></td>
<td></td>
<td></td>
<td>7.04</td>
<td></td>
<td></td>
<td>6.96</td>
</tr>
<tr>
<td>WANT to</td>
<td>2.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.75</td>
<td></td>
<td></td>
<td></td>
<td>2.98</td>
<td></td>
<td></td>
<td>3.04</td>
</tr>
<tr>
<td>Total</td>
<td>3.03</td>
<td>5.44</td>
<td>3.48</td>
<td>1.71</td>
<td>2.58</td>
<td>3.08</td>
<td>3.77</td>
<td>2.43</td>
<td>3.26</td>
<td>2.89</td>
<td>2.73</td>
<td>2.66</td>
<td>2.93</td>
</tr>
</tbody>
</table>
corpora used by Collins (2009), it might then be justified to claim that the corpus evidence here is not comparable to the well-known and already established data sets.

Table 5 shows that the ratio in ELF is 3.03 (2,000 in writing and 6,050 in speech), and this frequency falls well below the ratio in AmE (5.44) and just below AusE, but is larger than those of BrE (2.58) and NZE (1.71) in Collins (2009: 286). Most importantly, the results indicate that ELF usage is not as advanced as the inner circle varieties, but more advanced than the outer circle varieties. For SEA varieties the figure drops to 2.89 and the total outer circle variety ratio is 2.93. The results confirm that the ELF evidence presented in this section is not brought about by differences in the nature of corpus data that would be reflected in a skewed picture of modal and emergent modal frequencies.

5. Conclusions

This article has provided a new empirical, diachronically-informed angle to lingua franca uses of English. The approach is unarguably new, and there are possible caveats involved. As pointed out in Section 1, present-day ELF does not constitute a focused variety which has been coded in grammars, elaborated on and adopted in teaching or widely accepted by the population who resorts to English as an additional linguistic resource. As such, the situation resembles the state of affairs for the most part of the history of the English language in which focused varieties only emerge as standardized varieties in the modern period. Yet, the diachronic evolution of lexico-grammatical structures is well-documented in the native varieties even before the standard emerges. Therefore, the perspective adopted here, observing gradual developments and frequency drifts in lingua franca use, is motivated.

The results offer a multi-genre view to variability in ELF and contribute to various theoretical and empirical questions. Firstly, they expand the inventory of grammatical items which are studied in ELF (cf. Jenkins, Cogo & Dewey 2011: 289–290). I have illustrated one avenue for further research on how ELF speaker/writers adopt broader diachronic shifts in grammar. Secondly, they suggest that another sociolinguistic ization process, viz. globalization, should be subject to empirical studies in recent and ongoing grammatical change (cf. Leech et al. 2009: 236–270 discussing grammaticalization, colloquialization, Americanization, etc.). Thirdly, the results confirm that globalization had led to a situation in which the lines between various layers of English are getting increasingly blurred, and the diachronically-oriented approach does not support maintaining the division between ELF and indigenized L2 Englishes (cf. Section 1).
The results suggest that non-native use contributes to ongoing variability in ways that accelerate diffusion of grammatical change. The quantitative patterns described in Section 4 illustrate that ELF evidence does not surpass AmE as the leader of this particular change. Yet, they show that it falls in between AmE and BrE in terms of the diachronic fortunes of core and emergent modal frequencies in 4.1. A more detailed picture emerged in 4.2 in which the individual modal items were placed on a diachronic continuum. The results suggest that in cases of considerable quantitative changes in the recent history of the core varieties of English, the quantitative patterns in ELF polarize change. This means that many of the items which have undergone substantial recent increases exhibit substantially larger quantitative shifts in non-native use. Similarly, decreasing developments are further polarized, as seen with many outgoing core modal items. Lastly, the results in 4.3 suggest that an empirical perspective yields a more detailed view of global variability which does not confirm the canonized concentric model of varieties.

Future studies will shed light on (a) the qualitative patterns discernible in the individual core and emergent modals, and (b) to what extent these changes are regional and how they would be different across the ELF community.

References


Bowie, Jill, Wallis, Sean & Aarts, Bas. 2013. Contemporary change in the modal usage in spoken British English: mapping the impact of “genre”. In English Modality: Core, Periphery and Evidentiality, Juana I. Marin-Arrese, Marta Carretero, Jorge Arús Hita, & Johan van der Auwera (eds), 57–94. Berlin: Mouton de Gruyter.


CHAPTER 7

Building interdisciplinary bridges
MUCH: The Malmö University-Chalmers Corpus of Academic Writing as a Process

Anna Wärnsby,1 Asko Kauppinen,1 Andreas Eriksson,2 Maria Wiktorsson,1 Eckhard Bick3 and Leif-Jöran Olsson4
1Malmö University / 2Chalmers University of Technology / 3University of Southern Denmark / 4University of Gothenburg

This paper describes a corpus of writing as a process (MUCH), comprising English as a Foreign Language (EFL) student texts. The corpus will contain a large number of richly annotated papers in several versions from students of different performance levels. It will also include peer and instructor feedback, as well as tools for visualising the revision process, and for analysing the writing process and the peer and instructor feedback. MUCH will make it possible to study how texts develop and change in the course of the writing process and how feedback impacts the process.

Keywords: corpus, writing process, peer and instructor feedback, EFL

1. Introduction

Writing plays a crucial role at different levels in higher education. For researchers and PhD students, the pressure to publish seems to be increasing at a steady and unstoppable rate. Students, as well as universities, are evaluated on the basis of written products like Master’s and Bachelor’s theses. Many EFL students are expected to be confident users of both written and spoken English in their working life, as English is the main international corporate language. At the same time, students enter higher education with very different experiences of writing and with different levels of performance in English. This puts increasing pressure on universities to be able to cater for high levels of written performance both academically and professionally. It has, therefore, become more important to understand how writing at university can be taught and how to address writing-related issues of
student performance at various levels in higher education. This is not restricted to EFL contexts; the problem is two-sided as it does not only concern students, but also instructors and supervisors who need to know more about how to support students’ writing processes.

In the 1990s, language corpora – electronically stored and searchable text databases – grew to become a central tool in language research. Some of the great advantages of corpora are that research can be based on large data sets and that data can be made available to a wide range of users. However, the influence of corpora in the study of writing processes has been limited, partly because corpora typically contain only one version of a text. In this project, we aim to take an important step towards enabling the use of corpora to study writing as a process. Corpus-based research into the writing process is crucial as it provides, for example, new empirically grounded insights into the role of drafting, feedback, and revision, which in turn facilitates new pedagogical developments in the teaching of writing in order to support student performance in academic writing.

We aim to compile, systematise, and annotate a corpus of writing as a process: the Malmö University Chalmers Corpus of Academic Writing as a Process (MUCH). The corpus will be made searchable and will come with an interface that makes it possible to view differences between versions of texts and to connect changes to comments made by peers and instructors. What sets MUCH apart from other learner corpus projects is, first and foremost, its focus on writing as a process. This is realised primarily by the inclusion of several drafts of a paper, student self-reflective papers, and instructor and peer feedback in the corpus. The student papers in the corpus range from undergraduate to PhD levels, and the data is not limited to high-grade papers. MUCH caters for studies of various performance levels, genres and text types. In that way, it facilitates comparisons across proficiency levels, writing tasks and language backgrounds. Moreover, at the moment the data is being collected in EFL contexts, but will not be limited to these contexts in the future.

Another key issue is to expand corpus-based writing research in foreign and second language contexts, since a great deal of research on written feedback from teachers and students is based on studies carried out in first language contexts. The MUCH corpus aims to collect data produced in foreign and second language higher education contexts and hence to allow research on feedback in such contexts.

MUCH is a joint project between Malmö University, Chalmers University of Technology, the Swedish Language Bank at the University of Gothenburg (Språkbanken), and the University of Southern Denmark, in collaboration with The Language Archive at Max Planck Institute for Psycholinguistics.
2. Writing studies

At the heart of writing studies is naturally the act of writing. Writing is and has been central historically in “establishing an archive of thought, action and events for further social use” (Bazerman 2011: 8). In a modern literate society, the ability to understand and produce written texts cannot be overestimated. The field of writing studies takes its impetus from this understanding of the central role of writing and aims very practically to help people “use the written language more effectively, for both production and perception” (Bazerman 2011: 14). In this section, we describe approaches in writing studies that are relevant to this project and discuss the challenges and opportunities of using corpora in writing research.

2.1 Approaches in writing studies

Due to the practical aim of writing studies, the research questions in the field are often guided by pedagogical considerations. During the early phases of writing studies, text-types often related to the classical rhetoric modes of discourse – description, narration, exposition, and argumentation (Nelson & Kinneavy 2003) – were analysed for various features and rhetorical structures, which enabled the successful communication of the rhetorical purpose of each type. These text-types, along with corresponding rhetorical structures and features, were transmitted to students who were then asked to produce their own texts on the basis of these “good” models. Feedback was related to the success of composing texts with the desired structures and features (Galbraith & Rijlaarsdam 1999).

Partly as a reaction to product-oriented approaches to writing and writing pedagogy, process-oriented approaches emerged in the 1970s. Figure 1 illustrates such process-oriented approaches by emphasising the underlying cognitive processes of planning, researching, reflecting and revising, and foregrounds the different stages in the writing process itself, such as prewriting, drafting, reviewing, editing and proofreading.

Research on writing as a process thus concentrates on questions pertaining to how writers actively construct their ideas through the process of writing to fulfil communicative goals (cf., e.g., Flower & Hayes 1980). With this shift from product to process came various learning activities in writing pedagogy that aimed to support the writing process: outlining, drafting, revision (Galbraith & Rijlaarsdam 1999). These activities were pedagogically supported with feedback from both instructors and peers during the different steps of the process. The training of the ability to critique the work of another student, offer peer feedback or response, pedagogically contributed to individual student writer development.
However, at the same time as the cognitively-oriented process approach became more and more influential, more socially oriented approaches, as well as new types of product-oriented approaches, were gaining ground. Many of these approaches come together under the broad umbrella term of genre studies, often described as consisting of three main schools or approaches: the English for Specific Purposes (ESP) approach (Swales 1990), the Sydney School, based on the ideas of Halliday’s systemic functional linguistics (SFL) (Martin 1997), and the New Rhetoric approach (Miller 2005). Pedagogically, the ESP and SFL approaches have traditionally focused on uncovering conventionalised patterns of discourse with the aim of helping students understand how texts are written within a particular genre. By contrast, the New Rhetoric approach has emphasised the situational contexts and social purposes associated with different genres (Hyon 1996). There is still a tension between process and genre approaches to writing, even though several authors regard them as complementary rather than rivalry approaches (cf., e.g., Paltridge 2001). The MUCH corpus agrees with a complementary approach as its focus is on the writing process at the same time as it contains a variety of genres and text types (for example, research papers, abstracts, self-reflection, position papers).

Figure 1. The writing process approach (Curry & Hewings 2002: 34)
2.2 Corpora and their use in writing studies

During the 1970s and 1980s, the advent of more generally available computers and computer networks paved the way for a computer-based language research method known as corpus linguistics. It is a method of analysing language data using systematically compiled and electronically searchable collections of written or spoken language data. From a general linguistic perspective, this method contributed to richer descriptions of language usage within, for example, lexicography and grammar studies. Because they are based on much larger sets of actual usage data than was possible earlier, these descriptions are rich and help produce new, generalisable knowledge about language.

Within writing studies, corpus methods first allowed the easy comparison of textual features across different text types and registers. However, the contribution of language corpora to writing research has not been as extensive as was perhaps first expected (Ädel 2010; Flowerdew 2010; Gabrielatos 2005; Römer 2011). The scope of corpus-based research on writing has been limited by the design of the existing corpora as well as by the interests of researchers in the field of corpus research, as described below.

Many of the early writing corpora were collections of published texts written by experts; therefore, early large-scale descriptive studies on writing focused on such texts rather than on student writing (Alsop & Nesi 2009). This gap has later been addressed by several corpora collected to investigate the writing of university students specifically. Such corpora containing material produced by students are typically referred to as learner corpora, whether they consist of data from native or non-native speakers. Large-scale learner corpora have recently been collected in both first language environments (the British Academic Written English corpus (BAWE) and the Michigan Corpus of Upper-Level Student Papers (MICUSP)) and second language environments (The International Corpus of Learner English (ICLE) and The Varieties of English for Specific Purposes (VESPA)).

In addition, the majority of corpus studies on writing have concentrated on surface-level issues; that is, they have taken their starting point at the level of language and have not dealt extensively with higher-order concerns, such as text functions or argumentation (though there are notable exceptions; cf., e.g., Charles 2007 and Flowerdew 2008). For instance, studies have included aspects of vocabulary (Coxhead 2000), lexical bundles (Hyland 2008), and grammar (Hinkel 2004). The reason why most studies are of this type is that most existing corpora primarily lend themselves to language-oriented research. Ädel (2010) quite accurately

---

1. Here, we restrict our scope to corpora created specifically for the purpose of research on writing.
notes that one of the major challenges for the use of corpora in the teaching of writing is to broaden the scope and to allow for approaches that address higher-order concerns, such as rhetorical features and text functions.

Furthermore, both general and learner corpora such as BAWE and MICUSP, have so far only contained one version of a paper, typically the final one. From a process writing perspective (represented in Figure 1 above), a single draft or a final version of a paper can tell us very little about how the texts were produced. Berkenkotter and Huckin argue that “tracking and analysing the development of a report as it goes through various revisions yields unique insights about the epistemology of science” (1995: 49). By extension, we could argue that analysing report or paper revisions yields insights not only about the epistemology of science, but also about the formulations of knowledge in any disciplinary field, and how these formulations are transmitted in education. The single draft limitation does thus not only relate to the study of textual features but also to the way in which the text is developed as a result of critical thinking and argumentation. To extend the use of language corpora in writing research and pedagogy, it is, therefore, necessary to design new types of student writing corpora: corpora that have a clear process focus built into their design.

When designing new writing corpora, it is also necessary to address issues of proficiency level and scope (Krishnamurthy & Kosem 2007). For instance, both BAWE and MICUSP contain only papers awarded very high grades, that is, papers deemed by graders to be of high quality. From a descriptive perspective, these types of corpora permit studies of the finished product of high performance students, but they provide little information about the writing of a wider range of students. Pedagogically, these corpora may provide students with good models of writing; students are after all presumed to learn through observing good examples. Instead, this strategy works for some students, but writing instructors know that implementing good practice into one’s own writing is more complex, particularly when learning to write in a foreign language. It would be interesting to acquire more knowledge about the development of writing skills in students of various performance levels.

To further inform and improve writing pedagogy in general and writing process pedagogy in particular, the design of writing corpora should also allow for studies focusing on the relationship between written feedback and revision work. There are already several studies that have shown from various perspectives that feedback is effective for student learning (Nicol, Thomson & Breslin 2014; Cho, Schunn & Charney 2006; Gunersel et al. 2008). Still, however, recent reviews of feedback research (Hyland & Hyland 2006; Biber, Nekrasova & Horn 2011; Finnegan, Kauppinen & Wärnsby 2015) have argued that there is a lack of consensus concerning what constitutes effective feedback, partly as a result of different
interests and research designs. It is, therefore, not surprising that researchers across
different research traditions have argued for the urgent need of more research on
feedback. For instance, Nicol, Thomson and Breslin (2014) request studies on peer
response designs that are transferrable across disciplines and that address the issue
of feedback implementation. Similarly, Nelson and Schunn (2009: 396) see a need
for more research on the relationship between students’ understanding of received
comments and their implementation of the suggestions made. Moreover, feedback
categorisations need to be further refined (Anderson et al. 2010; Nelson & Schunn
2009; Eriksson et al. 2012; Hamer et al. 2014) by developing a more empirically
grounded network of feedback types. Hyland (2010), finally, calls for more studies
on the use of electronic tools to provide feedback. One of the difficulties in feed-
back research is, however, that data is not readily available to the wide research
community. It is, therefore, difficult to carry out some of the investigations that the
authors request. To conclude, more research on feedback is needed, and it should
be carried out on easily accessible data to support comparisons across contexts,
proficiency levels, and pedagogical setups. There is thus a demand to design a new
type of learner corpus to allow new types of research on student writing. Such
a corpus should contain multiple drafts of texts written in English by students
of various proficiency levels and from various backgrounds to facilitate studies
related to the writing process. It should be made available to researchers, teachers
and students, and it should be annotated to allow searches at various textual levels.

3. MUCH planned features

Annotation of grammatical features (tagging and parsing) is standard in major
corpora such as the Corpus of Contemporary American English (COCA) or the
International Corpus of British English (ICE-GB). Semi-automated grammatical
annotation of learner language has been performed on corpora such as ICLE and
BAWE (Part-of-speech [POS] tagging). However, automated tagging and parsing
is still a challenge, and tagging and parsing learner language is particularly chal-
lenging because it does not follow standard grammar rules consistently. The an-
notation tool that MUCH uses (EngGram parser; cf. section “Text Annotation”
for a fuller description) performs lemmatisation, POS-tagging, morphological
tagging and semantic parsing of types, roles and frames. It has been tested on
the MUCH pilot data and works well also with “dirty” learner data (i.e. data that
does not always comply with standard English grammar). Up to 90% of the data
in the pilot study were correctly annotated. The aim is also to further develop
the EngGram parser within the project to allow more complex structural analysis
of the MUCH data.
Previous research indicates that there is a need for a tool that provides better visualisations of the relationship between feedback and the implementation of feedback. For example, Nelson and Schunn (2009) used Track Changes in Microsoft Word to see what changes had been made between drafts. This is, however, a fairly rudimentary tool for the purpose. In the MUCH project, we will use the visualisation tool CollateX to track changes between different drafts of student papers. At the moment CollateX can be used to track textual differences between editions of text (see collatex.net). In this project, CollateX will be used to handle multiple drafts, but will also be developed to track comments across drafts. The adaptation of CollateX to MUCH will open new ways of analysing writing, revision and feedback processes.

3.1 What does MUCH contribute to the writing research community?

The corpus allows for new insights into the student writing process, the role of feedback and revision. The combination of grammatical annotation and multiple drafts makes it possible to combine top-down and bottom-up approaches to writing research. Corpus analysis is no longer synonymous with investigations based on one draft only. Below are some examples of research questions that can be explored on the basis of MUCH data.

Scholars interested in how critical thinking develops through drafting and revision can look at what choices the student makes in each draft to support her claims using evidence from secondary sources, what evidence the student selects, and how the student then qualifies her claims on the basis of this selection. Concretely, the researcher can analyse this by looking at the kind of sources the student includes, the logical connectors that are used, where the student chooses to paraphrase and where to quote, to mention but a few examples. Similar systematic, data-driven studies can be made, for example, on rhetorical strategies, development of argumentation, or metacognitive reflection.

Scholars interested in feedback can look at the type of feedback given by peers and instructors. Concretely, the researcher can classify what kind of peer and instructor feedback is offered and investigated, both quantitatively and qualitatively, and the effects of various feedback types: which feedback prompts revisions, where in the text it occurs, and when it occurs in the sequence of drafts. Similar systematic studies can be made, for example, on the use of metalanguage in feedback, the use of affirmative or corrective feedback, or comparisons between peer and instructor feedback. The MUCH data can also stimulate grounded scholarship, bridging the gap between academics and practitioners: instructors can study patterns and results of formative and summative assessment, analyse types and
times of peer and instructor intervention, identify strategies to scaffold learning for students at various levels of performance.

Scholars interested in stance taking (signalling one’s position on an issue through word choice) can look at the development of how claims are expressed, how evidence is described, and how the student then asserts credibility through these choices. Concretely, the researcher can analyse this by noting the prevalence of particular stance words such as modal verbs or adverbs in the different drafts, types of papers or levels of writing. Similar studies can be made, for example, on discourse and information structures, editing techniques, academic vocabulary development, grammar acquisition.

More broadly, writing skills are crucial for larger concerns of social cohesion, democracy and sustainability. West (2004: 7) identifies several “antidemocratic dogmas” related to knowledge communication threatening Western democracies: “We are losing the very value of dialogue – especially respectful communication – in the name of the sheer force of naked power.” For Duffy (2007: 89), the ability to dialogue is characteristic of “critical citizenry,” especially one composed of individuals unafraid to and capable of evaluating and critiquing institutions of authority. Fostering this type of critical citizenry is a fundamental feature of what West (2004) calls deep democracy.

Research on writing and writing pedagogy should thus aim to aid students’ acquisition of writing skills to achieve critical citizenry. Specifically, writing skills help students distinguish individual manifestations of a problem from systemic or root causes. Secondly, they help students assess rhetorical circumstances in the public sphere and intervene appropriately through writing and civic action. Thirdly, and perhaps more mundanely, they help students create purpose-driven documents for audiences beyond the classroom. Through learning how to write effectively and purposefully, students gain higher levels of rhetorical awareness, understanding of counterarguments, of how to tailor language to particular contexts and particular audiences, as well as an understanding of the complexity of arguments. Collecting data from writing courses also offers “a lens for exploring complexity and a vehicle for arriving at nuanced understandings of a lived reality that is inescapably characterised by ambiguities, shades of meaning, contradictions, and gaps” (Miller 2005: 169–70). Since MUCH will offer insight into the process of developing writing skills, it may indirectly offer insight also into the development of critical citizenry in diverse student populations and thus to social sustainability. In short, with MUCH as an open corpus resource, the research community gains valuable new tools and materials for data-driven interdisciplinary research on writing and writing pedagogy.
4. Building MUCH

The process of building MUCH follows the ethical guidelines established by
the Swedish Research Council (Vetenskapsrådet 2015), which are also in agree-
ment with the European Union ethical guidelines for researchers (European
Commission 2013). All contributing students sign a release-form, consenting to
their texts and metadata to be used for research and teaching. The following meta-
data is gathered: prior experience of writing courses, current course/programme
enrolment, age, gender, and mother tongue.

When we compiled the MUCH pilot, currently stored in the Language
Archive,\(^2\) we established data collection procedures, a functional archiving struc-
ture, and a better grasp of how feedback can be categorised. The MUCH pilot
consists of approximately 500,000 words in form of PDF-files of multiple drafts of
essays from various students from Malmö University (see Table 1 below).

Table 1. Papers included in the MUCH pilot

<table>
<thead>
<tr>
<th></th>
<th>Number of Students</th>
<th>Average Number of Field Research Texts</th>
<th>Average number of Argumentative Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autumn 2011</td>
<td>45</td>
<td>3.9</td>
<td>2.5</td>
</tr>
<tr>
<td>English Independent</td>
<td>11</td>
<td>3.7</td>
<td>2.2</td>
</tr>
<tr>
<td>European Studies</td>
<td>9</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Non-English Independent</td>
<td>5</td>
<td>2.4</td>
<td>1.4</td>
</tr>
<tr>
<td>Teacher Training</td>
<td>20</td>
<td>5.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Autumn 2012</td>
<td>57</td>
<td>5.1</td>
<td>3.4</td>
</tr>
<tr>
<td>English Independent</td>
<td>12</td>
<td>4.8</td>
<td>3.2</td>
</tr>
<tr>
<td>European Studies</td>
<td>18</td>
<td>4.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Teacher Training</td>
<td>27</td>
<td>5.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Spring 2012</td>
<td>33</td>
<td>4.2</td>
<td>2.8</td>
</tr>
<tr>
<td>English Independent</td>
<td>17</td>
<td>4.1</td>
<td>2.9</td>
</tr>
<tr>
<td>IMER</td>
<td>9</td>
<td>4.1</td>
<td>2.4</td>
</tr>
<tr>
<td>Non-English Independent</td>
<td>7</td>
<td>4.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Spring 2013</td>
<td>51</td>
<td>5.3</td>
<td>3.8</td>
</tr>
<tr>
<td>English Studies</td>
<td>41</td>
<td>5.3</td>
<td>4.0</td>
</tr>
<tr>
<td>IMER</td>
<td>10</td>
<td>5.3</td>
<td>2.7</td>
</tr>
<tr>
<td>Total</td>
<td>186</td>
<td>4.7</td>
<td>3.2</td>
</tr>
</tbody>
</table>

\(^2\) Crafoord Foundation, grant number 20120898.
As shown in Table 1, papers in the MUCH pilot were collected during 2011–2013 from a beginner academic writing course in which different cohorts of students – for example, those enrolled in the European Studies programme or those who took the course as an elective – are taught together. In the course, the students produced two types of papers: a field research report and an argumentative paper. The total number of contributing students in the MUCH pilot is 186. Only papers from those students who completed at least two drafts of each paper were collected. On average, almost five drafts per student of the field research report were included, and about three drafts per student of the argumentative paper. The selection processes and coding schemes for the student metadata, papers and instructor feedback on their papers were subsequently evaluated and adjusted for the future collection of MUCH proper. The data and metadata collected so far is stored at The Language Archive, is compatible with the search tools of the Archive and can be accessed upon request.

The following steps must be completed before the MUCH corpus can be extended and made available to the general research public:

1. data description
2. data collection and processing
3. text annotation
4. data availability and sustainability
5. software development

4.1 Data description

MUCH will include multiple drafts of student texts with peer and instructor feedback on these drafts. These texts comprise several different academic genres (bibliographic papers, argumentative papers, abstracts, scientific research papers, self-reflective papers). The estimated size of the corpus after a three-year collection period is approximately 6,500,000 words (including the MUCH pilot of 500,000 words). Peer and instructor feedback is not included in the word count as it is difficult to estimate how much feedback each draft elicits.

The estimated number of contributing students is about 600 (including the 186 students who have already contributed to the MUCH pilot). The contributing student population will be diverse: both graduates and undergraduates, from different disciplines, and with different levels of writing and language proficiency (determined by their achievements on the writing courses where the data is generated).
4.2 Data collection and processing

In the initial stage, texts will be collected from two different sources: from undergraduate and PhD academic writing courses at Malmö University and Chalmers, respectively. For each contributing student in these courses, we will collect metadata such as their prior experience of academic writing, age, gender, language background, and academic discipline. After collection, the papers and feedback will be anonymised; papers will be systematically renamed, and the relevant metadata will be linked to each paper draft through a code. Further, the peer and instructor feedback will be assessed to account for the scope and type of feedback.

4.3 Text annotation

The collected and processed documents are converted into xml-format to be automatically annotated using the English Constraint Grammar parser by Eckhard Bick at the University of Southern Denmark (EngGram: <http://visl.sdu.dk>) (Bick 2012). Annotation will be performed at morphological, syntactic and semantic levels, with searchable tags for parts of speech, lemmas, inflections, syntactic functions, dependency structures and semantic roles. The peer and instructor feedback will also be annotated for place (where it was given in the draft), scope (whether it is about a word, sentence, paragraph, section, or the whole paper) and type (correction, praise, criticism, suggestion).

All annotation layers will be converted from the EngGram parser format into xml-format, compatible with The Language Archive tools, such as ELAN. Originally an annotation tool, ELAN also contains a search interface module that allows for complex queries of xml-formatted language data. The generated xml-files will be stored in a corpus repository at The Language Archive, together with the MUCH pilot data. The repository makes it possible to adjust formats to the requirements of the search database. From the search database the data can then be made searchable through search interfaces such as the one included in CollateX.

4.4 Data sustainability and availability

Once the data has been processed and annotated, it will be archived in The Language Archive in two formats: the pdf-format retains the visual appearance of the type-written papers, the richly annotated xml-format allows complex queries to be put to the data through ELAN. All MUCH data, including the pilot data, will be in Open Format (pdf- and xml-files), and will use Open Source software tools: Arbil for creating the archive structure, LAMUS for uploading into the archive, ELAN for querying the data and CollateX for visualising feedback and changes.
between drafts. This contributes to the longevity of the data and access to it. The only non-Open Source tool is the EngGram parser, but an openly available application programming interface for its use will be set up in the course of the project. Open formats and open source software relieve MUCH from proprietary infrastructures and further ensure its sustainability.

The Language Archive as the corpus destination depository helps archive MUCH data sustainably and also allows for longevity of access. The Language Archive guarantees that data remain accessible for at least 50 years. As part of The Language Archive, the MUCH data becomes part of and accessible also via CLARIN (the Common Language Resource and Technology Infrastructure). CLARIN is a pan-European infrastructure which works to provide sustainable access to digital language data and advanced tools for analysing these data for scholars in the humanities and social sciences. All existing and future MUCH data and tools will conform to CLARIN standards to stimulate the sharing and comparison of data. MUCH will also have a website which will function as an access and information portal for the corpus project.

5. Concluding remarks

MUCH aims at building bridges between writing studies, linguistics and pedagogy. Upon completion, it will become a readily available resource for researchers, instructors and students alike. The research that will be made possible by the infrastructure will generate not only new insights into the writing process but can also contribute to the pedagogical development of writing in higher education.

References


CHAPTER 8

Discourse markers in L2 English
From classroom to naturalistic input

Gaëtanelle Gilquin
Université catholique de Louvain / FNRS

This chapter investigates how the context of acquisition, and more precisely the amount of naturalistic input received, may influence non-native speakers’ knowledge of English discourse markers. It considers three levels of analysis, from the more individual (foreign language learners having spent different periods of time in a target language country) to the more general (foreign language setting vs. official language setting), over an intermediate level of analysis comparing populations of foreign language learners from different countries. The corpus study carried out suggests that a higher degree of exposure to naturalistic language tends to have a positive impact on learners’ knowledge of discourse markers, resulting in more frequent use, better approximation of native speaker frequencies and, possibly, more fluent usage.

Keywords: target language input, acquisition context, discourse markers, English as a foreign language, English as a second language

1. Introduction

From a second language acquisition / foreign language teaching perspective, discourse markers (i.e. items such as you know, I mean or well) may not at first sight appear to be central elements of the language. They are generally said to contribute little propositional content to an utterance (Brinton 1996: 33–34), which limits their lexical value; their absence from interlanguage does not result in any identifiable error (Svartvik 1980: 171); and when present, they are often ‘tuned out’ by the hearer (Rieger 2003: 43). However, a more careful examination of these items in context reveals that they are as worthy of attention as any other items. While their propositional contribution is modest, they “contribute considerably at the pragmatic level, to the organisation of discourse (…), to marking a speaker’s attitude towards the propositional content of the utterance (…) and
to various functions pertaining to the speaker-hearer relation” (Andersen 2011: 599). Moreover, while their omission may not cause outright errors, a learner who does not use them may sound “dogmatic, impolite, boring, awkward to talk to” (Svartvik 1980: 171). Furthermore, while it is often the case that hearers tune them out, if used too frequently, discourse markers are “apt to be stigmatized (...) as a dysfluency” (Schourup 1985: 94). Thus, despite their trivial appearance, discourse markers serve crucial functions in language, and their absence or misuse may have unfortunate consequences for communication. It is therefore important to try to deepen our understanding of how they are acquired and used by learners, with the ultimate aim of facilitating their acquisition and improving their use.

The focus here will be on a variable that is likely to influence the acquisition and use of discourse markers, namely the context of acquisition, and more precisely the amount of naturalistic input learners receive. It is hypothesised that the more learners are exposed to this kind of input, the more they are apt to use discourse markers and to use them appropriately. This hypothesis is tested by means of three types of comparison: (i) a comparison between learners who have never been to an English-speaking country and learners who have spent periods of time in an English-speaking country; (ii) a comparison between foreign language learner populations representing different contexts of acquisition along the EFL (English as a foreign language) – ESL (English as a second language) continuum; and (iii) a comparison between foreign language learners and users of institutionalised second-language varieties. Each comparison involves speakers with different degrees of exposure to naturally-occurring language, which should contribute to attaining useful insights into the possible role of naturalistic input in the acquisition of discourse markers.

This chapter is organised as follows. Section 2 provides a brief overview of previous research on the use of discourse markers in learner English. Section 3 introduces the research questions of the study, while Section 4 describes the data and methodology used. Section 5 first offers some general results and then presents the outcomes of the three types of comparison aimed at determining the possible impact of naturalistic input. Finally, Section 6 discusses the results and proposes some concluding remarks.

2. Second language acquisition of discourse markers

Since the 1980s, a considerable body of research has been devoted to discourse markers. According to Fung and Carter (2007), research in the area has essentially focused on native language varieties (English predominantly, but also other languages). They note that “relatively limited research has been undertaken on the
range and variety of DMs [discourse markers] used in spoken English by second or foreign language speakers” (Fung & Carter 2007: 410). However, this situation is starting to change. With the growing availability of spoken corpora of learner language, especially for English, it has become easier to investigate typical features of spoken interlanguage, including pragmatic features. As pointed out by Overstreet (2012: 2), “[t]he study of pragmatic expressions owes its development directly to the availability of recording devices that allowed researchers not only to capture everyday spoken interaction, but also to transcribe it and investigate it ‘on the page/screen’ in ways that were almost impossible while the data was whizzing by ‘in the air’”. The tools and techniques of learner corpus research have facilitated the extraction and analysis of such expressions in interlanguage, which has contributed to a growing number of studies dealing with the subject matter.

With respect to the use of discourse markers by learners of English, several populations have been investigated. The mother tongue backgrounds which are represented include Chinese (e.g. Zhao 2013), German (e.g. Müller 2005), Spanish (e.g. Romero Trillo 2002) and Swedish (e.g. Aijmer 2011). Some studies focus on a single discourse marker (e.g. Li & Xiao 2012 on well), while others deal with a broader range of them (e.g. Aijmer 2004). What emerges from most of these studies is an overall underuse of discourse markers in learner English as compared to native English. However, since learners also appear to use a smaller range of discourse markers than native speakers, sometimes this results in an overuse of certain specific discourse markers. In Müller’s (2005) study, for example, all the discourse markers were found to be underused by German-speaking learners with the exception of well (see also Gilquin 2008 or Shimada 2014). More qualitative analyses often reveal a misuse of discourse markers as well. Aijmer (2011), for instance, shows that Swedish learners not only overuse well in comparison to native speakers, but also use it differently: they primarily use it for speech management functions (especially to stall for time) and tend to neglect its attitudinal functions (e.g. to mitigate disagreement).

This general underuse and misuse of discourse markers among learners of English could find an obvious explanation in the fact that these forms of language “are not explicitly taught in the classroom” (Hellermann & Vergun 2007: 158; see also Sankoff et al. 1997). Mukherjee and Rohrbach (2006: 216) note that, with the exception of well, “discourse markers are notoriously underrepresented even in modern [language-pedagogical] materials”. If discourse markers are not taught explicitly, the only way for non-native speakers to acquire them is incidentally, i.e. through exposure to language that includes (many) discourse markers. Such language is unlikely to be found in the language classroom setting. As suggested in Gilquin and Paquot (2008: 52), depending on the approach to language teaching that is adopted (e.g. communicative language teaching or grammar translation
method, teaching in the target language or mixture of target language and mother tongue), spoken input in the target language may be relatively limited. Moreover, this spoken input may not be particularly rich in discourse markers. First, oral pedagogical materials may be devoid of typically spoken features like hesitations and discourse markers (cf. Gilquin 2008: 143–144), either because these materials do not represent authentic speech or because they correspond to scripted registers such as broadcast news. Second, many teachers are not native speakers of the target language and hence their speech may exhibit the common interlanguage feature of not including many discourse markers. Alternatively, as pointed out by Hellermann and Vergun (2007: 176), the discourse markers found in teacher talk may essentially consist of “markers of discourse organization to show relationships between activities within the class period” (e.g. alright, okay, now), rather than markers “used to establish more local, interpersonal relationships in an interaction” (ibid.).

Accordingly, what seems to be needed in order to acquire discourse markers is a repeated exposure to naturalistic speech outside the classroom. However, since for most (EFL) learners classroom instruction constitutes the main access to the target language, they are unlikely to have been exposed to discourse markers and to have learned (how) to use them. As Aşik and Çephe (2013: 152) conclude in the context of Turkish learners, “[l]imited use of DMs [discourse markers] in spoken discourse reflects the unnatural context that EFL speakers are exposed to”. However, certain populations of non-native speakers fare better in this respect and thus stand a better chance of acquiring discourse markers. In Müller’s (2005) study referred to above, those learners who had benefited from more extensive contact with native speakers of English seemed to use certain discourse markers more frequently and in a more native-like manner than learners with little or no native speaker contact. Similarly, Hellermann and Vergun’s (2007) study of adult learners of English beginning their formal English language instruction in the US shows that “in general, the students who had been in the US the shortest period of time [before their enrolment in the language school] were less likely to have the pragmatic discourse markers as part of their language” (Hellermann & Vergun 2007: 167). In addition, preliminary results by Gilquin (2015a) suggest that users of institutionalised second-language varieties of English (i.e. varieties used in former British or American colonies where English has an official status, e.g. Hong Kong or Singapore) profit from a higher exposure to naturally-occurring English and produce more discourse markers than EFL learners.
3. Research questions

The research questions addressed in this chapter take as a starting point the above observation that discourse markers are not taught explicitly in the classroom and that the acquisition of these items must therefore be derived from naturalistic input in the target language. The general research question can be formulated as follows:

*Does increased exposure to naturalistic input facilitate the acquisition of discourse markers?*

This question will be considered at three different levels of analysis, each corresponding to a distinct sub-question:

a. *Do foreign language learners who have spent more time in an English-speaking country have a better command of discourse markers than foreign language learners who have spent less (or no) time in an English-speaking country?*

b. *Do foreign language learners who live in a more ESL-like environment have a better command of discourse markers than foreign language learners who live in a more EFL-like environment?*

c. *Do users of English as an institutionalised second-language variety have a better command of discourse markers than foreign language learners?*

It is hypothesised that, in each case, a higher degree of exposure will have a positive impact on speakers’ knowledge of discourse markers, given that naturalistic input represents learners’ best chance to acquire discourse markers. Similarly, the answer to the general research question is also expected to be affirmative, that is, learners with increased exposure to naturalistic input should demonstrate a better command of discourse markers. Frequency of use will serve as a rough approximation of the extent of learners’ knowledge of discourse markers (a higher number of discourse markers found in learners’ production suggests a more proficient command of discourse markers). In addition, what will also be taken into account is the proximity to native standards of frequency and the distinctiveness of the discourse markers for native English, as well as their appropriate and fluent use by learners (e.g. positioning within the utterance; see Section 5.5).

In addressing the three sub-questions above, we will progressively zoom out, from individual learners to populations of learners living in the same country, to types of learners living in different parts of the world but sharing the same general context of acquisition. The approach taken here will rely on the (by now generally accepted) assumption that EFL and ESL\(^1\) represent a continuum rather than a

\(^1\) In this chapter, I will exclude ESL situations in which immigrants learn the target language in a country where it is a majority native language.
dichotomy (e.g. Nayar 1997, Gilquin & Granger 2011) and that foreign language learners, who by definition mainly learn English through instruction in an educational setting, may receive more or less naturalistic input outside the classroom. By means of the three types of comparison necessary to answer the research questions (a) to (c) above, this chapter will seek to connect different varieties of English, produced by learners with various degrees of exposure to the target language and with different contexts of acquisition. In addition, it will attempt to build interdisciplinary bridges by bringing the fields of second language acquisition, which focuses on learner varieties, and contact linguistics, which includes the study of institutionalised second-language varieties, closer together. Since all the analyses carried out as part of this effort of rapprochement will be based on corpus data, the following section will briefly describe the data and methodology used in this study.

4. Data and methodology

The present study of the use of discourse markers by non-native speakers of English relies on the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010). LINDSEI contains informal interviews with 554 EFL learners (university students) from eleven L1 populations (Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish), amounting to a total of almost 800,000 words of learner language. The database as a whole was used to represent EFL production (Sections 5.1 and 5.4), but its different components were also taken into account in order to distinguish between L1 learner populations (Section 5.3) and between individual learners (Section 5.2). The metadata that are part of LINDSEI were examined to identify the learners’ mother tongue and to determine the time they had spent in an English-speaking country, if any. To provide a benchmark against which to evaluate learners’ use of discourse markers, data were also extracted from the Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004), an exact replica of LINDSEI with native speaker (British) interviewees, amounting to a total of approximately 120,000 words of native English. Finally, in an attempt to study the production of non-native speakers who, unlike the LINDSEI learners, are expected to receive regular naturalistic input, data from the International Corpus of English (ICE; Greenbaum 1996) representing institutionalised second-language varieties were analysed, viz. East African English, Hong Kong English, Indian English, Philippine English and Singapore English (Section 5.4).

A total of seven discourse markers were investigated, all of which are commonly cited in the literature, namely and so, and then, I mean, like, sort of, well and you know. Each of these lexical strings was searched for in the different corpora
and the hits were then manually filtered in order to retain only those occurrences that function as discourse markers. Among the sentences (1) to (3), for example, only (1) was kept as an instance of discourse marker; (2) was discarded on the grounds that it includes a prepositional use of like and (3) on the grounds that like is used as a verb.

(1) we play it like every day. four or five hours (LINDSEI-DU_025)
(2) it looked like two completely different worlds (LINDSEI-PL_046)
(3) you can see a lot of unusual and strange birds it’s something I like about New Zealand (LINDSEI-JP_038)

The remaining hits served as a basis for the calculation of the frequency of each discourse marker in the different corpora. In the case of LINDSEI, each occurrence was linked to the reference of the interview, as well as to the metadata associated with the interview, so that frequencies could also be calculated as a function of certain metadata. Relative frequencies are provided per 1,000 words or per 100,000 words (as indicated) depending on the size of the samples. For all frequency comparisons, statistical significance was calculated by means of the log-likelihood test and the computation of the probability (p)-value.

5. The use of discourse markers by learners with varying degrees of naturalistic input

This section presents the results of the different sets of comparisons corresponding to the three sub-questions listed in Section 3: a comparison between EFL learners with various lengths of stay in an English-speaking country, a comparison between learner populations living in more EFL- or ESL-like environments, and a comparison between foreign language learners and users of institutionalised second-language varieties. Before turning to these results, however, a broad overview of the frequency of discourse markers in learner English as compared to native English is provided.

2. The occurrences of and so, and then, I mean, sort of and you know in LINDSEI and LOCNEC were extracted and filtered within the frame of another study published in Gilquin and Granger (2015). All the remaining data were processed by myself.
5.1 Overall results

In order to gain a general idea of the use of discourse markers by learners of English, a comparison was drawn between LINDSEI and its native speaker counterpart LOCNEC (see Table 1). The results for all discourse markers taken together confirm the tendency identified by other researchers (see Section 2): discourse markers are generally underused by learners, with a relative frequency of 1,455 occurrences per 100,000 words in learner speech and a relative frequency of 2,679 occurrences per 100,000 words (almost twice as much) in native speech. According to the log-likelihood (LL) test, this difference is highly significant at the 0.0001 level.

<table>
<thead>
<tr>
<th></th>
<th>LINDSEI</th>
<th>LOCNEC</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>and so</td>
<td>73.47 (582)</td>
<td>64.19 (76)</td>
<td>n.s.</td>
</tr>
<tr>
<td>and then</td>
<td>170.42 (1350)</td>
<td>283.79 (336)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>I mean</td>
<td>147.20 (1166)</td>
<td>363.18 (430)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>like</td>
<td>156.29 (1238)</td>
<td>555.75 (658)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>sort of</td>
<td>51.88 (411)</td>
<td>474.67 (562)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>well</td>
<td>609.11 (4825)</td>
<td>435.82 (516)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>you know</td>
<td>246.80 (1955)</td>
<td>501.70 (594)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>ALL</td>
<td>1455.17 (11527)</td>
<td>2679.10 (3172)</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

Moreover, an examination of individual discourse markers reveals that, with the exception of and so, all the differences between LINDSEI and LOCNEC are statistically significant and mostly correspond to an underuse among learners, following the general trend emerging from the total results. Well is the only discourse marker that presents a significant overuse among learners, with 609 occurrences per 100,000 words in learner speech as against 436 in native speech. The underuses are all highly significant at the 0.0001 level, but a comparison of the relative frequencies in LINDSEI and LOCNEC shows that some differences are more marked than others. In particular, sort of appears to be over nine times as frequent in LOCNEC (475 occurrences per 100,000 words) as in LINDSEI (52 occurrences per 100,000 words), which points to a very heavy underuse of the discourse marker in learner English. Like is also quite heavily underused, being three and a half times less common among learners (156 occurrences per 100,000 words) than among native speakers (556 occurrences per 100,000 words). The remaining underused discourse markers, and then, you know and I mean, are between one and a
half times and two and a half times less frequent in LINDSEI than in LOCNEC. In terms of ranking, *well* comes first in LINDSEI and *like* fourth, whereas the opposite is the case for LOCNEC. *Sort of* is the least frequent discourse marker among learners while it is the third most frequent one among native speakers. *And then* comes third in LINDSEI and last but one in LOCNEC.

This overview reveals that, as expected from the lack of instruction in the use of discourse markers and the generally limited amount of naturalistic input in EFL environments (Section 2), learners tend to underuse discourse markers. The exceptions are *well*, which is significantly overused by learners, and *and so*, which does not display any significant difference between learners and native speakers. In what follows, I will attempt to identify those EFL learners who may have received greater exposure to naturalistic English and evaluate whether they may have derived any benefit from it in terms of usage of discourse markers.

### 5.2 Length of stay in a target-language country

All the learners who have contributed to LINDSEI are EFL learners, and as such, they are assumed to have received little exposure to English outside the classroom. However, among the 554 contributors, there are learners who have had the opportunity to travel to an English-speaking country and spend some time there – hence having been exposed to naturalistic input for a more or less extended period. The LINDSEI database includes information about any stays in English-speaking countries, indicating whether the learner spent time in an English-speaking country, and if so, for how long (cumulatively, that is, adding up the length of all his/her stays). About 40% of the interviewees reported no stay in an English-speaking country, while 45% reported a stay of some length; no information is available for some 15% of the interviewees. Among those learners who spent some time in an English-speaking country, the length of the stay varies from one week to nine years.

The possible impact of a stay in a target-language country on the use of discourse markers was investigated by associating each occurrence of one of the seven discourse markers in a given interview with the time (if any) the interviewee spent in an English-speaking country. For each period of time represented in the corpus, the relative frequency of the different discourse markers was calculated, making it possible to compare relative frequencies according to the time spent in an English-speaking country. If we first distinguish between the LINDSEI data produced by learners who never stayed in an English-speaking country (312,673 words) and those who spent some time in an English-speaking country (354,970 words), it appears, as displayed in Table 2, that there is a general, statistically significant increase for the latter group. While learners who never stayed in an English-speaking country show an overall relative frequency of 1171.51 discourse
markers per 100,000 words, the overall relative frequency for learners who spent some time in an English-speaking country amounts to 1625.77. Furthermore, Table 2 provides the results for each individual discourse marker. While the three discourse markers and so, I mean and you know do not display any significant difference between the two groups of learners, all other discourse markers present a significant increase. This seems to suggest that a stay in a target-language country usually results in a wider use of discourse markers, in line with the hypothesis advanced above (Section 3).3 Some discourse markers also appear to discriminate more reliably between learners who never went to an English-speaking country and those who did, in particular like and sort of, which display strikingly different relative frequencies for the two groups.

Table 2. Relative frequency per 100,000 words (and absolute frequency) of discourse markers in LINDSEI, according to whether learners stayed in an English-speaking country or not, with p-value of significance

<table>
<thead>
<tr>
<th>Discourse Marker</th>
<th>Stay = 0</th>
<th>Stay &gt; 0</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>and so</td>
<td>69.08 (216)</td>
<td>81.42 (289)</td>
<td>n.s.</td>
</tr>
<tr>
<td>and then</td>
<td>129.85 (406)</td>
<td>182.27 (647)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>I mean</td>
<td>153.20 (479)</td>
<td>163.11 (579)</td>
<td>n.s.</td>
</tr>
<tr>
<td>like</td>
<td>94.03 (294)</td>
<td>198.61 (705)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>sort of</td>
<td>13.43 (42)</td>
<td>86.77 (308)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>well</td>
<td>454.79 (1422)</td>
<td>663.72 (2356)</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>you know</td>
<td>257.14 (804)</td>
<td>249.88 (887)</td>
<td>n.s.</td>
</tr>
<tr>
<td>ALL</td>
<td>1171.51 (3663)</td>
<td>1625.77 (5771)</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

However, considering that the frequency of discourse markers may not only vary as a function of whether or not the learners went to an English-speaking country, but also as a function of how much time they spent there, a more precise analysis was carried out to determine the evolution in frequency of discourse markers.

3. Naturally, despite the statistically significant results, the possibility cannot be excluded that the difference between the two groups is linked to some other factor. It could for example be that learners who opt to spend time in an English-speaking country (e.g. through an Erasmus exchange) are those who, even before the stay abroad, are more motivated, more proficient, etc. A better approximation of the real impact of a stay in a target-language country would require the comparison of data produced before and after a stay abroad (cf. Gut 2012 on the LeaP corpus, which makes this possible by including data produced by thirteen learners of English and German before and after an extended stay in a target-language country; see also longitudinal studies such as Ren 2014 or Siyanova-Chanturia 2015 for a comparison of learners at the beginning and at the end of a study abroad period).
according to the length of the stay abroad. Figure 1 offers a graphical representation of the results, where the x-axis includes all monthly periods of time covered by the LINDSEI data and the y-axis shows the relative frequencies per 1,000 words. A (linear) trendline revealing the overall direction of the LINDSEI relative frequencies is provided, as well as an indication of the average relative frequency in native English (LOCNEC). The graph in Figure 1a displays the results for all the seven discourse markers taken together. It exhibits an upward trendline which suggests that, overall, the frequency of discourse markers tends to increase with a longer period of time spent in an English-speaking country. There are five main peaks in frequency: at 4 months, 15 months, 18 months, 27 months and 48 months. However, it is only from the 15th month onwards that the frequency in learner speech starts to equal (and even exceed) the frequency in native speech.

![Graph](https://via.placeholder.com/150)

**a.** Results for all the discourse markers

![Graph](https://via.placeholder.com/150)

**b.** Results for the discourse marker *and so*
c. Results for the discourse marker *and then*

d. Results for the discourse marker *I mean*

e. Results for the discourse marker *like*
Discourse markers in L2 English

f. Results for the discourse marker *sort of*

![sort_of](image)

LINDSEI  LOCNEC  Trendline LINDSEI

well

![well](image)

LINDSEI  LOCNEC  Trendline LINDSEI

g. Results for the discourse marker *well*

![you_know](image)

LINDSEI  LOCNEC  Trendline LINDSEI

h. Results for the discourse marker *you know*

Figure 1. Relative frequency per 1,000 words of discourse markers in LINDSEI, according to length of stay in an English-speaking country (in months), with LINDSEI linear trendline and LOCNEC average relative frequency.
Among individual discourse markers, most display an upward trendline, similarly to the overall results. However, two exhibit a tendency to decrease with increased time spent in an English-speaking country: and so and well. Concerning well, it should be noted that it is generally overused by learners (Section 5.1), so that a downward trendline actually corresponds to a tendency towards more native-like usage. This is further indicated by the fact that the LINDSEI trendline in the graph for well in Figure 1g gradually comes closer to the native average frequency. With respect to and so, it is the only discourse marker which does not present any statistically significant difference in frequency between learner and native English (Section 5.1) and it is also the least frequent discourse marker in native English. Accordingly, here again, a decrease in frequency does not necessarily point to a deficit on the part of the learners. The graph for and so in Figure 1b reveals that the LINDSEI trendline is slightly above the native speaker average for shorter periods of time spent in an English-speaking country and slightly below for longer periods, but the trendline is very close to the native average throughout. By contrast, the other discourse markers and then, I mean, like, sort of and you know all have an upward trendline (sometimes minimally), which gradually comes closer to the native speaker average, with the trendline for sort of even exceeding it towards the end. The evolution of frequencies is more or less regular, depending on the discourse marker. For instance, I mean (Figure 1d) presents many ups and downs, whereas sort of (Figure 1f) shows a very slow and more gradual increase, with only a few peaks in frequency for some of the longest stays abroad.

Previous studies have shown that for some aspects of language, even a very short stay abroad may make a difference. Llanes & Muñoz (2009), for example, demonstrate that stays of three to four weeks make it possible to improve listening comprehension and oral fluency significantly. By contrast, for other aspects a much longer stay might be necessary before any noticeable impact on learners’ language production is felt. Thus, according to Siyanova and Schmitt (2007), it is only after a stay abroad of more than twelve months that learners appear less likely to opt for a one-word verb (e.g. cancel) rather than a multi-word equivalent (e.g. call off). In an attempt to identify the threshold at which a stay in an English-speaking country has the most significant influence on the use of discourse markers, the relative frequencies of each discourse marker before and after a certain threshold were compared and the statistical significance of the difference was computed. The thresholds correspond to each period of time abroad (in months) represented in the LINDSEI database: less than one month vs. one month and more, less than two months vs. two months and more, less than three months vs. three months and more, etc. Considering the frequency of the seven discourse markers as an aggregate, it turns out that the most highly significant difference, corresponding to an increase, is found between learners who stayed
less than four months in an English-speaking country and those who stayed four months or more (LL = 265.18, \( p < 0.0001 \)). In other words, it seems as if, globally, a four-month stay in an English-speaking country is likely to have the strongest impact on the frequency of use of discourse markers. There can be increases in the use of discourse markers before the end of this time period, but these increases are less significant. The same threshold emerges from the comparisons for *and then* (LL = 48.60, \( p < 0.0001 \)) and *I mean* (LL = 67.73, \( p < 0.0001 \)), which suggests that four months is a critical period in the acquisition of discourse markers. For *like*, *sort of* and *you know*, the required time abroad is longer: the increase in the frequency of *like* is most significant after ten months in an English-speaking country (LL = 268.44, \( p < 0.0001 \)), while for *sort of* it is eighteen months (LL = 330.68, \( p < 0.0001 \)) and for *you know* forty-eight months (LL = 18.38, \( p < 0.0001 \)).\(^4\) In comparison with *and then* and *I mean*, it thus appears that *like*, *sort of* and *you know* require more time abroad and hence more exposure to naturalistic language in order to display a maximally significant increase in frequency. With respect to *well*, the most significant increase actually occurs after only one week in an English-speaking country (LL = 130.01, \( p < 0.0001 \)). However, since, as mentioned above, *well* tends to be overused in interlanguage, this increase does not coincide with a threshold that should be reached by learners in order to become more native-like. Such a threshold is to be found in a (slightly) significant decrease in the frequency of the discourse marker, which occurs after the longest possible stay in an English-speaking country in the data, namely nine years (LL = 5.5, \( p < 0.05 \)). This seems to suggest that learners pick up *well* extremely rapidly when abroad, but need a very long and presumably intensive exposure to naturalistic English before they realise that this discourse marker is not so frequent (at least in the register represented in LINDSEI and LOCNEC) and should not be used too often. Finally, the most significant result for *and so* (LL = 8.87, \( p < 0.005 \)) corresponds to a decrease in frequency between learners with less than eight months in an English-speaking country and those with eight months and more. This result is consistent with the downward trendline highlighted above and, like *well*, it seems to point to an adjustment after some time abroad when learners start to realise that the discourse marker is not as frequent as they might have imagined (though the amount of time is shorter here than in the case of *well*).

\(^4\) Because not all periods of time abroad are represented in LINDSEI, it might be that some thresholds are actually reached earlier. In the case of *you know*, for example, no learners with a stay abroad of 37 to 47 months are included in the database, so that the same impact as that observed for the 48-month period might have been achieved after, say, a 38-month or 43-month period; the 48-month threshold for *you know* is thus partly arbitrary.
This section has examined individual learners and considered the possible impact of a stay in an English-speaking country on their use of discourse markers. The following section deals with populations of learners who share the same mother tongue and live in the same country.

5.3 EFL-like vs. ESL-like environment

LINDSEI includes data produced by EFL learners from different mother tongue backgrounds and who live in different countries. In total, eleven populations are represented in the currently released version of LINDSEI: Bulgarian (BG), Chinese (CH), (Belgian) Dutch (DU), (Belgian) French (FR), German (GE), Greek (GR), Italian (IT), Japanese (JP), Polish (PL), Spanish (SP) and Swedish (SW). The general context of acquisition of English may differ for these populations, and in particular, some of them can be claimed to live in a country that exhibits some characteristics of an ESL environment. For example, while Sweden is/was typically classified as an ‘EFL country’ (cf. Strevens 1992: 36), it is widely recognised that “already in the early 1990s a shift from EFL to ESL was seen under way in the Nordic countries” (Pahta & Taavitsainen 2011: 609). Sweden is also the top European country when it comes to knowledge of English, with 89% of the population being able to carry out a conversation in English according to the 2006 survey on Europeans and their Languages (ibid.).

For each population represented in LINDSEI, the compilers of the national subcorpus provided information about the learning context of the interviewees. On the basis of this information, which is summarised in Gilquin et al. (2010: 47–56), it was possible to determine whether the population was situated more towards the EFL end of the EFL-ESL continuum (see Section 3) or more towards the ESL end. Table 3 shows which LINDSEI populations had access to English through written media (newspapers, magazines, books), spoken media (television, cinema, radio) and the Internet. Using these three criteria, four distinct groups can be identified:

5. More components are in preparation, representing Arabic, Basque, Brazilian Portuguese, Czech, Finnish, Lithuanian, Norwegian, Taiwanese and Turkish populations of EFL learners.

6. It should be emphasised that the features describe the situation as it was at the time of the data collection. Thus, the French data were collected between 1995 and 1997, at a time when there was practically no Internet access; nowadays the Internet offers increased exposure to English in French-speaking Belgium (Gilquin et al. 2010: 50). Similarly, while today there are many undubbed English television programmes in Germany, this was not the case in 2004 when the German data were collected (ibid.).
a. Chinese and French-speaking LINDSEI populations, for whom English newspapers, magazines and books were not easily available, almost all English-speaking TV programmes were dubbed, and no access to the Internet (in English) was possible.

b. German, Italian, Japanese and Spanish LINDSEI populations, for whom almost all English-speaking TV programmes were dubbed, as for the preceding group, but for whom English newspapers, magazines and books were widely available and who in addition had access to English through the Internet.

c. Bulgarian and Polish LINDSEI populations, who could easily buy English newspapers, magazines and books, and also had access to the Internet in English, and whose exposure to spoken media in English was existent but moderate: in Bulgaria, “several English and American TV programmes were broadcast without any translation or subtitles” (Gilquin et al. 2010: 47–48); in Poland, all radio shows were in Polish and all English-speaking TV programmes were dubbed, but movies in cinemas were subtitled.

d. Dutch-speaking, Greek and Swedish LINDSEI populations, who had access to English through written and spoken media, as well as through the Internet.

Table 3. LINDSEI populations’ access to English

<table>
<thead>
<tr>
<th>Written media</th>
<th>Spoken media</th>
<th>Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-FR</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>GE-IT-JP-SP</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>BG-PL</td>
<td>✓</td>
<td>✗/✓</td>
</tr>
<tr>
<td>DU-GR-SW</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The eleven populations of LINDSEI can thus be arranged as follows along the EFL-ESL continuum, with the populations to the left presenting a relatively typical EFL situation, with hardly any exposure to naturalistic English, and the populations to the right presenting a more ESL-like situation, with quite some exposure to English in more natural contexts:

+EFL-like

CH-FR < GE-IT-JP-SP < BG-PL < DU-GR-SW

+ESL-like

Figure 2 shows the relative frequencies per 100,000 words of the seven discourse markers (taken together and individually) for the different L1 populations represented in LINDSEI. The populations are ordered according to the above EFL-ESL continuum and an indication of the relative frequency in native English (LOCNEC; cf. green bar) is provided. The total frequencies (Figure 2a) form the
following scale, from the population with the lowest frequency of discourse markers (Japanese) to the populations with the highest frequency of discourse markers (French, Dutch, Swedish and Polish), where the arrow signals a statistically significant difference between the two adjacent populations:

\[ \text{JP} < \text{GR-CH} < \text{IT} < \text{GE} < \text{SP-BG} < \text{FR-DU-SW-PL} \]

This scale does not exactly coincide with the EFL-ESL continuum established above. However, it is noticeable that in both cases the Japanese, Chinese, Italian and German populations are situated towards the lower end of the scale, whereas the Bulgarian, Polish, Dutch and Swedish populations are situated towards the upper end. Thus, for these populations, there seems to be some correspondence between the (supposed) degree of exposure to naturalistic English and the frequency of use of discourse markers. The Dutch and Swedish learners, in particular, who are expected to receive the most input in English, are among the populations producing the most discourse markers, as predicted by the hypothesis advanced above (Section 3). The Spanish learners tend to use slightly more discourse markers than one would expect given their degree of exposure to English, but it is mainly the French and Greek populations who behave unexpectedly. The French-speaking learners in LINDSEI are assumed to have had little access to naturalistic English but present some of the highest frequencies of discourse markers, while the Greek learners, who are supposed to have received more exposure to naturalistic English, do not use discourse markers very often. As will be shown below, the high frequency of discourse markers in the French component of LINDSEI is mainly due to a substantial overuse of \textit{well}. This discourse marker was described earlier as rather atypical of native English and its use is therefore unlikely to result from large amounts of exposure to English. As for the Greek learners, it might be that their exposure to English in everyday life is not as high as suggested by Table 3. In this respect, it is interesting to note that while an ESL-like environment is often acknowledged for Swedish (see above) and Dutch learners (see e.g. McArthur (2003) or Edwards (2016) on Dutch learners from the Netherlands, whose situation is arguably quite similar to that of Belgian Dutch-speaking learners), this is not usually the case for Greek learners. In addition, the Greek learners represented in LINDSEI are remarkable in that none of them spent any time in an English-speaking country, which suggests that the general context of acquisition and the learners’ individual situations (Section 5.2) might in combination account for some of the results.
a. Results for all the discourse markers

b. Results for the discourse marker *and so*

c. Results for the discourse marker *and then*
d. Results for the discourse marker *I mean*

![Graph showing the frequency of *I mean* in different languages.]

**like**

![Graph showing the frequency of *like* in different languages.]

e. Results for the discourse marker *like*

**sort of**

![Graph showing the frequency of *sort of* in different languages.]

f. Results for the discourse marker *sort of*
Discourse markers in L2 English

Results for the discourse marker well

Results for the discourse marker you know

Figure 2. Relative frequency per 100,000 words of discourse markers in LINDSEI per L1 population, with reference native frequency in LOCNEC (green bar)

Furthermore, considering the other graphs in Figure 2, which show the frequencies of the various discourse markers per L1 population, a great deal of variation can be observed. For example, the Italian learners use the discourse marker and so much more often than the other populations (including the native speakers); by contrast, the discourse marker like is used most often by the Polish, Spanish and Swedish learners, while the Italian learners hardly use it all. Since and so and well have been shown not to be particularly typical of native speaker usage (and so is relatively uncommon in native English; well is overused by learners and outnumbered by sort of, you know and like among native speakers), a positive correlation

7. A reviewer points out that the Italian equivalent of and so is very frequent in Italian. This suggests another possible explanation for some of the results outlined in this section, namely crosslinguistic influence. Identifying cases of crosslinguistic influence would require further analysis (including a frequency analysis of the equivalents of the discourse markers in the learners’ mother tongue) and will therefore not be attempted here.
between their frequency in learner English and the amount of input that the learners are assumed to have received cannot really be expected. Apart from the very high frequency of \textit{and so} among Italian learners (cf. Figure 2b), the discourse marker has another peak in the French component of LINDSEI. The other populations use it with a lower frequency, corresponding more or less to that in native English. \textit{Well} (Figure 2g) is most massively overused by the French-speaking learners, but the Bulgarian and Dutch-speaking learners use it almost as often. The Polish, Spanish, Swedish and German learners all overuse the discourse marker, though to a lesser extent. Among those learners who overuse \textit{well}, two correspond to learner populations that occupy the ESL-like end of the continuum (see above), namely the Dutch-speaking and Swedish learners. This confirms the point made in Section 5.2 with respect to stays in English-speaking countries: large amounts of input seem to be necessary before learners realise that a given discourse marker should not be used too often – larger amounts, probably, than most foreign language learners can receive, even in ESL-like environments. The remaining discourse markers are all characterised by a general underuse in learner English (cf. Table 1) and thus can be expected to increase in frequency with a higher exposure to naturalistic English. Accordingly, the discourse marker \textit{and then} (Figure 2c) is most frequently found among the Dutch-speaking and Swedish learners, the two populations that are most ESL-like. By contrast, the Japanese learners present a very low frequency of the discourse marker, as one would expect given the limited exposure to English they receive. However, they use \textit{and then} more often than the Greek learners, who come last in terms of frequency of the discourse marker – despite their supposedly input-rich environment. This result for the Greek learners mirrors the general trend of discourse markers in Greek EFL production (see above). The other results for \textit{and then} do not correspond very closely to the EFL-ESL continuum either, with e.g. the Chinese learners displaying a relatively high frequency of the discourse marker and the Polish learners a relatively low one. Similarly, the frequencies for \textit{I mean}, \textit{like}, \textit{sort of} and \textit{you know} according to L1 population show both convergences and divergences with the EFL-ESL continuum. With respect to \textit{I mean} (Figure 2d), the Swedish and Bulgarian learners seem to benefit from their higher exposure to English. These two groups use the discourse marker more often than any other learner populations, including the Japanese learners, who use it very rarely, in keeping with their more limited access to naturalistic English. By contrast, the Dutch-speaking, Polish and Greek learners display a low frequency of \textit{I mean} and the Chinese learners a high one, which contradicts the predictions made on the basis of the learners’ linguistic environment. The discourse marker \textit{you know} (Figure 2h) turns out to be most frequent among the Polish learners (who use it more often than the native speakers) and the Greek learners, as could be expected from Table 3. However, it is more frequent...
than expected among the Spanish learners, who use it slightly more often than the Swedish and Dutch-speaking learners, that is, two populations that have been described as more ESL-like. As for *like* (Figure 2e) and *sort of* (Figure 2f), these discourse markers are interesting in that they are only common among certain populations and are hardly ever used by other populations. This is especially striking for *sort of*, which is common among the Swedish learners only, thus suggesting that these learners have taken advantage of their input-rich environment to acquire this discourse marker, unlike most other L1 populations. Surprisingly, while the Dutch-speaking learners are also assumed to live in an input-rich environment, they use *sort of* much less often (although they come second after the Swedish learners). There are more populations who use *like* regularly, and among those several are situated towards the upper end of the EFL-ESL continuum: the Polish, Swedish and Dutch-speaking learners. The Spanish and German learners also appear to use *like* regularly, despite their more restricted access to English in everyday life. However, because *like* and *sort of* also tend to be used very rarely by several populations who are assumed to receive limited exposure to naturalistic English (e.g. Chinese or Italian learners), it could be that these two discourse markers are more difficult to acquire than others and therefore discriminate more reliably between learners who have relatively easy access to English in natural contexts and those who mainly receive their input from a classroom context. It will be reminded that this was also suggested by the results of the comparison between learners who spent time in an English-speaking country and those who did not (Section 5.2).

The above analysis reveals that there is no exact correspondence between the linguistic environment in which learner populations acquire the target language and the frequency with which they use discourse markers typical of native English. However, some tendencies emerge which show that certain populations with facilitated access to naturalistic English use certain discourse markers more often than other populations whose access to naturalistic English is more limited. The following section considers a more global context of acquisition, with the aim of testing the link between naturalistic input and use of discourse markers from yet another perspective.

5.4 Foreign vs. institutionalised second-language varieties of English

Among non-native varieties of English, one can distinguish foreign language varieties, such as those represented in LINDSEI (EFL), and institutionalised second-language varieties of English (ESL).8 The latter varieties are used in countries,
normally former British or American colonies (e.g. India or the Philippines), in which English functions as an official or semi-official language, serving various administrative, legal, social, educational as well as literary purposes. The consequence of this (semi-)official status of English is that people in these countries are exposed to the target language more often than EFL learners and in a wider range of contexts, including the natural contexts of everyday life. Table 4 summarises some of the differences between EFL and ESL situations. In both cases students take English language courses at school or university, but in ESL countries other school/university subjects are taught in English as well, so that ESL learners receive more exposure to English in an educational context. Moreover, they receive more exposure to English in their everyday setting, for example through the media or through interactions with the administration. By contrast, EFL learners receive limited input in the target language outside an educational setting, at least for intra-national communication. Another difference is that English is culturally foreign in EFL situations, whereas in ESL situations it has historical roots due to colonial history.

Table 4. Comparison of EFL and ESL situations

<table>
<thead>
<tr>
<th></th>
<th>EFL</th>
<th>ESL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure in educational setting</td>
<td>Instruction in the target language</td>
<td>Instruction in the target language</td>
</tr>
<tr>
<td></td>
<td>- Instruction through the target language</td>
<td>- Instruction through the target language</td>
</tr>
<tr>
<td>Exposure in natural setting</td>
<td>Limited or non-existent, except for international communication</td>
<td>In a variety of situations, especially for institutional communication</td>
</tr>
<tr>
<td>Cultural/historical context</td>
<td>Culturally foreign</td>
<td>Historical roots</td>
</tr>
</tbody>
</table>

In line with the hypothesis detailed in Section 3, one can expect discourse markers to be more widespread in ESL than in EFL, given the higher degree of exposure to English in ESL situations. A further reason which points to a more widespread use in ESL is to be found in the fact that discourse markers are claimed to be signs of acculturation (cf. Hellermann & Vergun 2007). Such acculturation is more likely to occur in an ESL environment where English is culturally anchored than in an EFL environment where it is culturally foreign. These two elements

Sweden and the Netherlands resembling ESL countries more and more (cf. Section 5.3), and ESL countries like Hong Kong leaning increasingly towards the EFL part of the continuum. The features in Table 4 therefore do not cover the whole spectrum of situations but merely describe the more typical EFL and ESL situations.
can be assumed to reinforce each other in facilitating the production of discourse markers in ESL situations.

In order to test this prediction, the above results for EFL (all LINDSEI components taken together, see Section 5.1) were compared to results for ESL, using data from the International Corpus of English (ICE). Five ESL populations were represented in the samples used (East African, Hong Kong, Indian, Philippine and Singaporean), amounting to a total of a little over one million words of ESL. To make up the corpus, henceforth referred to as ICE-L2, only the ‘direct conversations’ sections of ICE were selected, corresponding to relatively informal and spontaneous speech, similar to the LINDSEI interviews. In an attempt to minimise the possible effect of register differences on the results, frequencies of the discourse markers in LINDSEI and ICE-L2 were not compared directly. Instead, frequencies were compared indirectly through the yardstick of a native (ENL; English as a Native Language) corpus – LOCNEC for LINDSEI, and ICE-GB, the 180,000-word British component of ICE, for the ICE-L2 data. Since LOCNEC and ICE-GB are built according to the same design criteria as LINDSEI and ICE-L2 respectively, the comparability of these pairs of corpora is guaranteed. However, differences between LINDSEI/LOCNEC on the one hand and ICE-L2/ICE-GB on the other should be treated with caution as they may be due to differences between the two registers (cf. Fuller 2003 on the difference between interviews and casual conversations in the use of discourse markers).9

The results of the comparisons are represented visually in Figure 3. Overall (Figure 3a), discourse markers are more common in interviews (LINDSEI/LOCNEC) than in conversations (ICE-L2/ICE-GB), but in both registers non-native speakers appear to underuse discourse markers compared to native speakers (the difference is highly significant, with $p < 0.0001$). However, the ratio between non-native speech and native speech (NNS/NS) is higher in ESL than in EFL (66% vs. 54%), which shows that the ESL speakers are proportionally closer to their native peers (ICE-GB) than the EFL speakers are to theirs (LOCNEC). Nevertheless, the other graphs in Figure 3 reveal that not all discourse markers follow this trend. And so (Figure 3b) is not a particularly common discourse marker, regardless of register. As discussed in Section 5.1, for and so there is no statistically significant difference between LINDSEI and LOCNEC. Similarly, the difference between ICE-L2 and ICE-GB is not statistically significant. By contrast, the remaining discourse markers all display a highly significant difference between

---

9. Although this will not be dealt with here, it should be underlined that the figures for the ESL speakers actually hide some variation between different populations (East African, Hong Kong, Indian, Philippine and Singaporean speakers) and between individual speakers, as was shown to be the case among the EFL speakers.
native and non-native speech in either register \( (p < 0.0001) \). And then (Figure 3c), like (Figure 3e) and you know (Figure 3h) present an identical profile in being underused by the EFL speakers and overused by the ESL speakers. The behaviour of both non-native groups is thus somehow problematic, but in a different way. This difference confirms the prediction made in Section 3 that, with their higher exposure to naturalistic English, ESL speakers are more likely to use discourse markers than EFL speakers. It should also be emphasised that, according to the NNS/NS ratio, the ESL speakers are closer to their native peers than the EFL speakers are. This suggests that in addition to acquiring discourse markers more easily than EFL speakers, ESL speakers might also be storing more accurate information about their frequency in certain registers. I mean (Figure 3d) and sort of (Figure 3f) differ in frequency between interviews and conversations but are significantly underused by non-native speakers in both registers. Here again, however, the NNS/NS ratio reveals a proportionally larger gap between LINDSEI and LOCNEC than between ICE-L2 and ICE-GB, thus signalling a better approximation of native frequencies for the ESL speakers than for the EFL speakers, despite the underuse characteristic of both groups. Well (Figure 3g), finally, is remarkable in that it is overused by the EFL speakers (see Section 5.1) and underused by the ESL speakers, which seems to contradict the hypothesis according to which more exposure to the target language should lead to an increased use of discourse markers. Also, the NNS/NS ratio shows that this time it is the EFL speakers who, proportionally, come closer to their native peers. It was pointed out earlier that the massive use of well in EFL is not a particularly desirable characteristic since well is not so common in LOCNEC compared to the other discourse markers. The EFL speakers could thus be claimed to (over)use well to the detriment of other discourse markers, hence diverging from the native speakers who show a more even distribution of the seven discourse markers. The situation for the ESL speakers is different because well is the most frequent discourse marker in ICE-GB (which, incidentally, supports Fuller’s (2003) finding that well is more typical of conversations than of interviews), so that ESL speakers would be fully justified in using this discourse marker much more frequently. In Section 6, I will argue that this unexpected underuse of well in ICE-L2 may be due to competition from indigenous discourse particles that could serve as alternatives to well.
a. Results for all the discourse markers

b. Results for the discourse marker *and so*

c. Results for the discourse marker *and then*
d. Results for the discourse marker *I mean*

![Graph showing results for *I mean*]

e. Results for the discourse marker *like*

![Graph showing results for *like*]

f. Results for the discourse marker *sort of*

![Graph showing results for *sort of*]
The above analysis suggests that, overall, both groups of non-native speakers underuse discourse markers in comparison to native speakers. However, disregarding the differences in register, which are also reflected in native English, it can be said that the ESL speakers tend to use most discourse markers more often than the EFL speakers and to better approximate the frequency of discourse markers of their native peers – as predicted by their higher exposure to naturalistic English. In the next section, I will reconsider some of the results of the preceding sections in a more qualitative light by examining how discourse markers are used in context and whether they contribute to fluency or not.

5.5 A more qualitative peek

In the preceding sections, it has been assumed that more exposure to naturalistic English should lead to an increased use of discourse markers. However, one could also expect this higher exposure to result in better usage of discourse markers, i.e. use in registers where discourse markers are more likely to appear, in positions...
where they will not hinder communication, with meanings and functions that correspond to native usage, etc. While a comprehensive qualitative analysis of all discourse markers considered in this study is clearly beyond the scope of this chapter, in what follows, two examples will be used to briefly illustrate that, besides frequency, the behaviour of discourse markers in context should also be considered for a thorough evaluation of the possible impact of naturalistic input on the 'nativeness' of the output.

The focus here will be on the position of discourse markers within the utterance. A distinction will be drawn between uses that do not interrupt a structure (as in *It seems very expensive I mean our five star is nearly a hundred pounds*, where both what precedes and what follows the discourse marker constitute complete structures) and uses that do interrupt a phrase or a closely-knit structure (as in *it's a bit you know flattering*). Using this classification, Gilquin and Granger (2015) compared the use of *you know* by French-speaking and Polish EFL learners, with native speaker data as a baseline. As has been discussed in Section 5.3 (cf. Figure 2h), while the French-speaking learners display a relatively modest frequency of *you know*, as could be expected given their limited exposure to naturalistic English, the Polish learners present the highest frequency of the discourse marker, using it much more often than the Swedish, Greek and Dutch learners, who are assumed to receive the largest amount of input, and even more often than the native speakers themselves. However, from a qualitative point of view, it turns out that the French-speaking learners use *you know* in a manner more similar to that of the native speakers than the Polish learners do. Compared to 35% of interrupted structures in native English, the French-speaking learners show 37% and the Polish learners 60%. Polish learners’ use of *you know* is therefore more disfluent, occurring between words that should normally be uttered together – for example between a copula and the nominal part of the predicate (e.g. *he stopped being you know humorous*) or between a preposition and the rest of the prepositional phrase (e.g. *that was some guy from you know the upper classes*). This shows that quantitative and qualitative behaviour do not always go together, and that a high frequency of discourse markers does not necessarily imply that these discourse markers are used appropriately and in a way that contributes to fluency. With respect to the research question concerning the impact of naturalistic input on the use of discourse markers, this specific example suggests that the somewhat surprising position of the Polish learners, at the top of the list in terms of frequency

10. This comparison relied on 150 occurrences of *you know* randomly selected from the same set of data used here to represent the French-speaking and Polish EFL populations; the native occurrences were similarly extracted from LOCNEC.
of *you know* but not in terms of degree of exposure, might be counterbalanced by a less than perfect command of the discourse marker.

The same type of analysis was performed on EFL and ESL data for the discourse marker *I mean*, on the basis of 200 occurrences per variety, randomly selected from LINDSEI and ICE-L2 respectively. The native baseline consisted of 200 randomly selected occurrences of the discourse marker in LOCNEC and ICE-GB. The results (cf. Figure 4) show that cases of interrupted structures are most common in EFL, where they represent 32%, and least common in native English, where they represent 17%. ESL occupies the middle ground, being closer to native English than to EFL with 23%. This finding is consistent with the prediction that ESL speakers will have a better command of discourse markers as compared to EFL speakers (see Section 3); it also confirms the frequency-based results of Section 5.4, which indicated a general tendency for discourse markers in ESL (including *I mean*) to be closer to the native baseline than discourse markers in EFL. Examples (4) and (5) illustrate the interrupted use of *I mean* in EFL. They show that foreign learners, possibly due to their input-poor environment, often insert *I mean* in positions where it is rather disruptive, causing the message to become difficult to understand and disfluent as a result. While such examples can also be found in ESL, cf. (6), they are less widespread than in EFL, as one would expect given ESL speakers’ higher exposure to naturalistic (and presumably generally fluent) discourse in English.

(4) I had the possibility to: to understand a lot about *I mean* different type of culture cultures  
(LINDSEI-IT_032)

(5) they will *I mean* .. broadcast many good films in at night around nine-thirty every day  
(LINDSEI-CH_013)

(6) I am not talking of *I mean* the traditional grammar but at least the modern approaches to grammar should at least be given to the students  
(ICE-IND_S1a-078)

This type (and other types) of qualitative analysis should be undertaken for the remaining discourse markers and applied to all the comparisons carried out in Section 5, including the comparison between learners with different lengths of stay in an English-speaking country. Only then would it be possible to make reliable claims about the relationship between the frequency of discourse markers and appropriateness of usage. However, the preliminary results above suggest that also from a qualitative point of view, degree of exposure can have an impact on the use of discourse markers and closeness to the native standard.

Having progressed through the three levels of analysis for which the possible impact of exposure to the target language can be investigated and having
considered the data from a more qualitative perspective, I now turn to a discussion of the results and some concluding remarks.

6. Discussion and conclusion

The above analysis has established some degree of correspondence between the amount of naturalistic target language input that non-native speakers are exposed to and their command of discourse markers: non-native speakers with more exposure to naturalistic English are more likely to use discourse markers (and to use them appropriately). The answer to the general research question (Section 3) thus seems to be affirmative: increased exposure to naturalistic input arguably facilitates the acquisition of discourse markers. This is true at all the three levels of analysis carried out here (cf. three sub-questions), but slightly more apparent at the level of the individual learners (time spent in a target language country) and at the level of foreign language vs. official language setting.

As is evident from the analysis, however, the correspondence between degree of exposure and command of discourse markers is not perfect, but merely represents a tendency (similar to the trendlines in Figure 1). There are several reasons that could explain this lack of perfect correspondence and the sometimes ‘messy’ results. The first relates to the apparently idiosyncratic use of discourse markers. Analyses at the level of individual speakers, including native speakers, suggest that the use of discourse markers may vary quite strikingly from one speaker to another. For instance, a given speaker may show a very strong preference for you know and another one rely almost exclusively on I mean. If preferences for discourse
markers are so personal, then the results of a corpus-based study of discourse markers might to a certain extent be biased by the composition of the corpus. A second reason is that learner language is a highly heterogeneous variety that can be affected by many variables, some of which might explain, perhaps better than the ‘input’ variable, part of the variation observed in the analysis. To give but one example, proficiency is not necessarily dependent on the degree of exposure to naturalistic input (some learners have a very good general command of the English language despite receiving little input outside the classroom) and could account for certain unexpected results. This could for instance justify the position of the Greek learners at the top of the EFL-ESL scale in terms of exposure and at the bottom of the list in terms of frequency of discourse markers: as it so happens, 80% of the Greek learners in LINDSEI evaluated according to the Common European Framework of Reference for Languages had a B2 level or lower, whereas all the evaluated Dutch-speaking and Swedish learners (the other two populations at the top of the EFL-ESL scale, but usually also at the top of the frequency list) had a C1 or C2 level (see Gilquin et al. 2010: 11). A third factor that could explain why certain (groups of) learners do not use many discourse markers despite an input-rich context of acquisition is that easy access to the target language does not necessarily mean that learners are actually exposed to it and benefit from this exposure. EFL learners may have stayed in an English-speaking country for several months but spent most of their time alone or with speakers of their native language; learners in an ESL-like environment may deliberately opt out of any contact with the target language by not buying newspapers or magazines in English, by avoiding TV programmes in English, etc. In other cases, this lack of contact may not be deliberate but may have to do with the particular circumstances of the learner, e.g. a learner living in a rural area and not being able to benefit from the same exposure as learners living in urban centres. Put simply, just because exposure to the target language is theoretically possible does not mean that learners actually are exposed to it. Also, the language they are exposed to may not meet the standards of native English as represented in LOCNEC or ICE-GB. This is true of exposure in the classroom (see Section 2), but also of exposure in natural settings. For example, the British or American series that Dutch-speaking and Swedish learners watch on television constitute oral input, but being scripted and enacted, their language may have characteristics that are different from authentic, spontaneous language. As for users of institutionalised second-language varieties of English, they are likely to receive most of their input in English from other users of institutionalised second-language varieties who are also non-native speakers of English. In this respect, it must be pointed out that sometimes (native) English discourse markers may have to compete against local norms and local (i.e. L1/indigenous) discourse markers (cf. Gilquin 2015a: 114–116). In Singapore English, for instance, the particle lah
appears to be more frequent than most native English discourse markers (Gilquin 2015a: 115). Since *lah*, in some of its functions, seems to bear certain similarities with *well,* it might be that the unexpected underuse of *well* in ESL (Section 5.4) is in fact due to some of its functions being fulfilled by the indigenous particle *lah.*

Finally, there is the thorny issue of register. In the present analysis, comparisons have been drawn with native speaker corpora representing the same register as the non-native speaker corpora (interviews for LINDSEI/LOCNEC and conversations for ICE-L2/ICE-GB) and cases of under-/overuse have been established on that basis. However, this relies on the assumption that non-native speakers are aware of register differences and of how discourse markers behave in each distinct register. But in reality, non-native speakers may not have register-specific models in mind that they could try to imitate. They may have a unique native speaker model that is based on a single register (e.g. the most typical register or the register to which they have been exposed most often) or a general model that brings together characteristics of different registers. The fact that in native English *well* turns out to be more frequent in conversations than in interviews (Section 5.4), for instance, could explain why even learners with extended stays in English-speaking countries continue to overuse it in the LINDSEI interviews (Section 5.2) if they rely on conversational models to assess the frequency of the discourse marker.

From a more methodological point of view, this chapter has combined the traditional learner corpus approach with less common practices in the field, notably the use of metadata beyond the over-researched variable of the mother tongue and the related individual analysis considering each learner separately in order to determine his/her context of acquisition (length of stay abroad). Both of these practices come with their own disadvantages. For one, data tend to be more difficult to interpret when more individual variation is taken into account. Further, despite the reasonably large size of LINDSEI, the relevant samples become quite small when certain variables have been selected (e.g. samples corresponding to very long stays abroad), thus making the claims that are based on these samples less reliable (as pointed out in footnote 4, some of the thresholds of time abroad identified as being crucial in the acquisition of discourse markers may be partly arbitrary because not all periods of time abroad are covered in the LINDSEI data). However, these practices have also made it possible to gain new insights into the nature of interlanguage and the influence of learners’ environment and experiences abroad.

---

11. For example, *lah* and *well* arguably share the function of “editing marker for self-correction” (Svartvik 1980: 175 on *well*; see Low & Deterding 2003: 64 on *lah*). Both markers also seem to have a softening function, notably when introducing a “dispreferred response” (Hellermann & Vergun 2007: 160 on *well*; see Wong 2004 on *lah*). In fact, in Goddard (1994: 148) *well* appears as a translation of *lah* in the example *Masuk lah! Aku hantar engkau!* (‘*Well, get in. I’ll give you a lift!*’).
on their language production. Both the individual and sociolinguistic approach could be taken even further, for example by considering the frequency of each discourse marker in the production of each individual learner or by relying on more detailed learner profiles providing information about every aspect of exposure to the target language. Such information may include measurement of input received through reading, listening to music, using social networking websites, etc. (cf. Gilquin 2015b: 30–31), similarly to the approach taken by Sankoff et al. (1997) in their study of the use of French discourse markers by Anglophone Montreal French speakers. Through its comparison of EFL and ESL, the present chapter has also contributed to the collaborative effort, started a few years ago, to bring second language acquisition and contact linguistics closer together (cf. Nesselhauf 2009, Mukherjee & Hundt 2011, Gilquin 2015a). It has shown that despite the obvious impact of different contexts of acquisition, with lower or higher degrees of exposure to the target language, foreign and institutionalised second-language varieties share some features in their use of discourse markers, such as a general tendency to underuse them and a habit to insert them within closely-knit structures, where they produce a disfluent effect. It is to be hoped that this and other interdisciplinary bridges will continue to be built and that they will lead to the discovery of new areas of research.

References


CHAPTER 9

Processing of aspectual meanings by non-native and native English speakers during narrative comprehension

Andreas Schramm and Michael C. Mensink
Hamline University, St. Paul, Minnesota / University of Wisconsin-Stout, Menomonie, Wisconsin

Languages have unique systems of language forms, meanings, and conventions for expressing narratives and temporal structure, grammatical aspect and its meaning and use being part of it. In second-language acquisition, little is known what temporal concepts and concurrent forms are comprehended at certain development stages and whether resulting mental representations are similar between native and non-native adult English speakers. In this study, we investigate whether readers attend to semantic content and draw causal inferences. Advanced non-native, unlike native, readers appear to not notice aspectual meanings and, apparently, the input is not cognitively registered; implicit learning of aspect seems unlikely. In native readers, aspect affects the availability of situations enabling causal inferencing, and imperfective aspect appears to be mentally stored in-focus.

Keywords: grammatical and lexical aspect, aspectual meanings, language acquisition, narrative processes, text comprehension

1. Introduction

Language users’ personal experiences are generally represented in a non-linear fashion. However, narratives must take a linear sequential form, when language users render their experiences in a text. For example, we can recount an accident between our car, another car, and a deer in multiple orders: our car tried to avoid the deer subsequently hitting the other car and the deer; in passing another car, our car hit a deer and the other car; a deer appeared suddenly while our car passed another car, and we hit both; etc. To be able to deviate from such imposed linguistic sequencing, languages provide means to encode situations in a narrative
through some form of temporal structure. The linguistic means take many shapes, such as ordering of events, lexical items, morphosyntactic markers, and inherent semantic content of situations. When we cross cultures and languages, the need to communicate experiences comes with us, but may have to take a different form. Each language has a unique system of language forms, meanings, and conventions for expressing narratives and also specifically temporal structure.

While in the field of second-language acquisition research much is known about the acquisition\(^1\) of temporal structure expressed through morphosyntax and the lexicon, much less information is available about the meanings and concepts involved in expressing temporal structure. There is an even greater lack of empirical data on how learners comprehend the concepts involved in denoting time in language to create an appropriate mental representation of the situations in a text. Several empirical studies have addressed the development\(^2\) of linguistic means to express temporal information, but few, if any, have investigated what temporal concepts are comprehended in narratives at a particular developmental stage and whether the resulting situation models (mental representations of text events) resemble those of native speakers. The present study explores whether the concepts connected with two morphosyntactic aspects (Progressive and Simple Past) create the same narrative connections of different strengths in an advanced-speaker group’s interlanguage as they create in native speakers’ situation models. Relevant interlanguage development and instructional issues, the temporal structures expressed through Progressive and Simple Past, and native-speakers’ cognitive processing\(^3\) of these two aspects will be described in the following sub-sections.

2. Second language acquisition

Within the last three decades, educators have generally expressed the view that, with the almost exclusive influence of meaning-based Communicative Language Teaching, attention to grammar and form has been eradicated (Nassaji 1999, 2000; Xiao-xia 2006). The field is therefore now opting towards a blended approach

---

1. While “acquisition” is often used to refer to naturalistic learning and in opposition to formal methods of “language learning”, we follow the common practice of using these two terms interchangeably (Davies 2005).

2. We use “development” following Vygotsky as “a collaborative process in which individuals move from what they are incapable of to what they are able to do […]” (Lantolf 2012).

3. Processing is “the cognitive activities that underlie different aspects of language learning and use (e.g., retrieving or recognizing the meaning of a word or producing a sentence)” (Gass & Mackey 2012: 596).
the focus-on-form approach (Hama & Leow 2010; Long & Robinson 1998; Slabakova 2014; Williams 2005). Recent literature suggests that grammar-oriented Form-Focused Instruction, the focus-on-form approach, and Communicative Language Teaching need not be mutually exclusive (e.g., Nassaji 1999, 2000). In fact, most teachers value the practice of blending these approaches because it elicits, enhances, and combines the grammatical and sociolinguistic aspects of language (Genesee 1987; Lightbown 1992; Long 1991; Nassaji 1999; Swain 1985). Many empirical studies have suggested the same (e.g., Allen, Swain, Harley, & Cummins 1990; de Graaff 1997; DeKeyser 1994, 1995; Harley & Swain 1984; Nassaji 1999, 2000; Robinson 1996; Swain 1995).

The combined approach is captured by Long’s (e.g., 1983, 1996) Interaction Hypothesis stating that, for interlanguage (a learner’s language systems that develop during acquisition) development, interaction in addition to a focus on form is necessary. Focus on form relies on explicit information in classroom second language acquisition to draw learners’ attention to and notice targeted linguistic forms (e.g., Leow 2001; VanPatten & Cardierno 1993). As several theorists state (Hedge 2000; Long 1988; Skehan 1996; Tarone & Yule 1989; Willis 1996), a sole and strict focus on communicative competence can suffocate the need for linguistic accuracy and potentially stunt the learner’s interlanguage system. A strictly communicative approach does not ensure that learners notice (pay conscious attention) to the language they are producing. Attending consciously to linguistic features is a strategy proven quite effective for language learners (Bardovi-Harlig & Reynolds 1995; Egi 2004; Leow 2001; Lopez-Ortega & Salaberry 1998; Schmidt 1990). Similarly, Ellis et al. put forth the position that both conscious and unconscious processes play a role in every learning episode and cognitive task (Ellis, Lowen, & Erlam 2006), and Schmidt (1990, 1993, 1995) states that language learning requires both intentionality and incidental learning. While the findings from studies of learners who have not received explicit instruction on language forms have not been consistent, researchers have postulated that some attention to form is required to facilitate acquisition of second or foreign language (Leow 1997, 2000; Rosa & Leow 2004; Rosa & O’Neill 1999; Sachs & Suh 2007; Williams 2004, 2005). There is, however, disagreement on what cognitive resources have to be assigned (Leow 1997, 2000; Tomlin & Villa 1994; VanPatten 1990, 1993, 1997; Williams 2004).

Furthermore, the cognitive status of data during encoding in working memory (WM) needs further investigation (Egi 2004; Williams 2005): Whether data that is “noticed” (attended to consciously) and stored (Schmidt 2001; Robinson 1995), becomes accessible intake (is registered cognitively through further processing (Robinson, Mackey, Gass & Schmidt 2012), such as during pragmatic retrieval.
Andreas Schramm and Michael C. Mensink

from long-term memory (LTM\(^4\)), and can be integrated into a learner’s knowledge base in later stages of the acquisition process (Carroll 1999; Gass 1997; Robinson 2003). There has been a call that this cognitive status should be explored further (Hama & Leow 2010). One variable in uncovering processing behavior in WM has been the impact of task (Leow 2001; Rapp & Mensink 2011). For example, a problem-solving task appears to facilitate noticing as opposed to a holistic reading task (Leow 2001). One such task commonly investigated are narratives and, specifically in it, the semantic domain of aspect, because aspect is difficult to acquire for second language learners (Bardovi-Harlig 2000; Andersen & Shirai 1996).

The further investigation of aspect in narratives is in line with the claim that language should primarily be considered a resource for meaning rather than a system of rules for using forms (Mohan & Beckett 2001). The use of linguistic forms varies with respect to communicative task (von Stutterheim & Klein 1989). It has to be redefined, and thus relearned, within each new context while their meaning stays the same. The constant linguistic features across tasks are therefore semantic meanings, not forms. This is recognized in the “concept-oriented or meaning-oriented approach,” a lesser-used approach to the study of language acquisition that investigates meanings (or concepts) and the forms that are associated with it (Bardovi-Harlig 2014, 2000; Klein 1995). It also studies the relationship between linguistic forms and contextual information (Klein 1995). The approach looks at how a concept, such as past time or aspect, is expressed and how means of expression for a concept change over time. For example, in the earliest stage of tense-aspect acquisition, the pragmatic stage (Meisel 1987), a learner relies on extralinguistic resources (e.g., conversational turns), or universal ordering principles of events (e.g., the iconicity assumption (Dahl 1984) or principle of natural order (Klein 1987)). Next, in the lexical stage, learners employ time adverbials and connectives (e.g., and then). In the final stage, the morphological stage, time content is expressed using the verbal tense-aspect system alongside adverbials. Most of these investigations into meaning acquisition have been focused on language production; the current study extends this research into comprehension.

Additionally, in an effort to understand successful language acquisition better, comparative study of adult second-language learners has led to some preliminary understanding of how language processing in language learners differs from that of mature native speakers (Clahsen & Felser 2006). These processing studies of very high-proficiency learners have shown that learners rely on morphological-semantic information during processing in the same way as native speakers, but less

---

4. WM and LTM are defined as a combination of time and function. WM is the memory in use immediately during reading, while LTM contains the mental representation that ensues after the immediate processing during reading has ceased.
so on syntactic cues. However, more studies of less proficient advanced learners comparing their second-language morpho-semantic and grammatical development with processing in native speaker are needed. Finally, applied linguists (e.g., Andersen & Shirai 1996; Ayoun & Salaberry 2008; Bardovi-Harlig 2000) and cognitive psychologists (e.g., Becker, Ferretti & Madden-Lombardi 2013; Magliano & Schleich 2000; Mozuraitis, Chambers & Daneman 2013) have mostly used linguistic explanations that blend temporal and non-temporal information to investigate the semantic domain of aspect rather than focusing on temporal information only. These theories and an alternate explanation of aspect exclusively focusing on temporal meanings will therefore be summarized in the following section.

3. Linguistic theory of aspect

The semantic domain of aspect has been subdivided into two components: situation aspect (also referred to as Aktionsart or lexical aspect) and viewpoint aspect (also called grammatical or morphosyntactic aspect; Comrie 1976; Klein 1994; Smith 1997). Situation aspect is expressed through semantic and lexico-syntactic means and understood compositionally on the sentence level. It has been classified into categories, or event notions (Schopf 1984), which are commonly called states, activities, accomplishments, and achievements (Ryle 1949; Vendler 1957). States are expressed by predicates like be musical; activities by predicates like walk in the park; accomplishments by predicates like write a book; and achievements by predicates like win the race. Since this widely used categorization does, however, rely on non-temporal as well as temporal criteria (Klein 1994), Klein’s categorization of situation aspect into 0-, 1-, and 2-state situations employing purely temporal criteria is used in this study instead. Only 2-state situations are used in the current investigation of causal inferences.

Klein’s classification is based on how many time periods, or states, are expressed as part of a situation’s inherent temporal content. 0-state situations largely correspond to what is typically called “co-extensive states or processes,” 1-state situations correspond to “events,” “processes,” or “activities,” and 2-state situations to telic (with an endpoint) “accomplishments” and “achievements.” 2-state situations consist of two lexicalized time periods, for example the period before and after the book is written in the predicate write a book, and thus are durative as well as telic. Furthermore, the second state “projects its lexical properties onto the post-time” (Klein 1994: 105). This allows for a sharper focus on the temporal basis for the causal inferences drawn by speakers. Our design thus differs from Mozuraitis et al. (2013) and Magliano and Schleich (2000) in terms of the situation aspect employed. Their studies used Vendler’s (1957) durative-versus-punctual-situation
distinction and selected only durative situations (e.g., walk) excluding telic punctual achievement and accomplishment situations (e.g., win the race, write a book).

The two periods become visible differently for example when 2-state situations are put in the Past Progressive or the Simple Past. In the sentence Stephanie was writing a book only part of the first period is denoted. The test question If Stephanie was writing a book and was interrupted while writing a book, has Stephanie written a book yet? must be answered in the negative (Schopf 1984). The second lexical period denoting that the tire is changed is not expressed in the Progressive, only in the Simple Past along with part of the first period (Klein 1994). By contrast, 1-state situations do not have such a contrast. The sentence Bob was dancing with Nancy (used by Mozuraitis et al. (2013) and Magliano and Schleich (2000)) elicits a positive answer in response to the question If Bob was dancing with Nancy and was interrupted while dancing with Nancy, has he danced with Nancy yet? 2-state situations have different semantic temporal content in focus when they are expressed in the Progressive compared to 1-state situations (Klein 1994; Schopf 1984).

This difference plays out in working memory as follows. Like Magliano and Schleich (2000), and unlike Mozuraitis et al. (2013), Schramm (1998) found in two studies that lexical content from critical sentences in either the Simple Past or Past Progressive was equally activated in working memory immediately after the critical sentence. Based on the theoretical explanations discussed above (Klein 1994; Schopf, 1984) that only part of the first state in 2-state situations is expressed in the Progressive, the study tested whether this expressed semantic content either stays active in WM or is tagged as being in focus in LTM. To this end, the design used in the current study was developed, and 4–7 sentences were inserted between a cause and a surprise effect to guard against recency (close proximity) effects immediately after the cause (cf. Table 1 below). Results showed that working memory activation of lexical targets from either Progressive or Simple-Past sentences had subsided 4–6 sentences later, which is in line with Magliano and Schleich (2000). Yet, semantic content from the first time period of the situation in the Progressive was still relatively more accessible in WM for causal inferencing after a causal coherence break that was introduced at this point in the narrative and thus must have been tagged as “in focus” compared to Simple-Past content (Schramm 1998). In the story in Table 1 below, a truck was more accessible in the Progressive version of the sentence. Semantic content in readers’ LTM for both the long and short texts was also checked. Here, too, Progressive semantic content from the first time period of 2-state situations was relatively more accessible. The effect was stronger in short than in long texts. The in-focus tag of Progressive semantic content appears to weaken, as mental representations that contain features corresponding to the situation model are updated with in-coming new text information.
Table 1. Sample Narrative

<table>
<thead>
<tr>
<th>Narrative text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Shelley was in a hurry.</td>
</tr>
<tr>
<td>2. a. She was <em>passing</em> a truck.</td>
</tr>
<tr>
<td>b. She <em>passed</em> a truck.</td>
</tr>
<tr>
<td>3. Suddenly she had to swerve to the right.</td>
</tr>
<tr>
<td>4. <em>Within</em> split seconds, <em>several</em> thoughts <em>raced</em> through her head.</td>
</tr>
<tr>
<td>5. <em>The</em> car belonged to the <em>company</em> her dad works for.</td>
</tr>
<tr>
<td>6. Would she get in trouble for taking it without asking?</td>
</tr>
<tr>
<td>7. Would he punish her by not letting her use his car again?</td>
</tr>
<tr>
<td>8. She felt a strong bump.</td>
</tr>
</tbody>
</table>

Overall, Progressive aspect used in conjunction with 2-state situations thus appears to have a stronger cataphoric effect in terms of foreshadowing relevance of semantic content for future narrative events in WM and the situation model than it may when used in conjunction with other types of situation aspect where target phrases from situations in the Progressive were relatively unavailable in WM four sentences post the critical aspectual sentence. A cursory analysis of the situation aspects in the study by Magliano & Schleich reveals that they are fairly evenly split between 1-state and ambiguous 1- or 2-state situations on the one hand and 2-state situations on the other. The semantic content of the poststate of roughly half of the situations in the Progressive may therefore not be lexicalized and relatively inaccessible. This would account for some of the differences in activation patterns in WM between the two studies. A comparison of the cataphoric effect in LTM was not possible because of the differences in study design.

Viewpoint aspect is expressed through morphosyntactic forms and their meanings conveyed in the predicate (Comrie 1976; Klein 1994; Schopf 1984; Smith 1997). The two main viewpoint aspect categories in English are the perfective *vb-ed* or *have+vb-ed* and imperfective *be+vb-ing* aspect. The current study contrasts perfective Simple Past and imperfective Past Progressive. The former is typically characterized as completed, the latter as ongoing. When 2-state situations are in the imperfective (ongoing), “the time for which an assertion is made falls entirely within the time of the situation” (Klein 1994: 108). In the case of (2a) *Shelley was passing a truck* in Table 1 above, the narrated time fully lies within the process of passing a truck. Her swerving to the right could very well lead to an accident with the truck.

The pragmatic function of morphosyntactic aspect in determining temporal and specifically causal relations between situations in narratives has been explored elsewhere (Hopper 1982). The function of aspect can be described as one of ‘discourse sequencing.’ Narratives contain foregrounded situations in the perfective and backgrounded situations in the imperfective. Foregrounded situations
succeed one another in the narrative in the same order as in the real world. Backgrounded situations are concurrent with foregrounded situations and do not succeed each other or the foregrounded situations. Grammatical aspect has been empirically found to influence causal inferencing by native speakers (NS) between situations in a narrative text (Schramm 2001, 1998). Since causality appears to be a more effective force in organizing situations in readers’ mental representations of narratives than time, space, or character (Taylor & Tversky 1997; Tzeng & van den Broek 1997; Zwaan, Magliano & Graesser 1995), we concentrate on causal inferencing in our current study of the role that aspect plays in narrative comprehension and in second language development. As for non-native speakers (NNS), perfective aspect and discourse foregrounding evolve first according to the Aspect and Interlanguage Discourse Hypotheses (Bardovi-Harlig 2000) and are initially associated with 2-state situations. Later, this is followed by imperfective aspect and backgrounding, which is connected with 1-state situations. Below we will describe processing of aspect by NS.

4. Cognitive processing of aspect

There has been a marked increase in interest in the effect of aspect on discourse processing in native English speakers. Aspect affects the construction of a coherent mental representation, which is the basis of comprehension, and contains features corresponding to the situation model readers build on five dimensions (causation, space, time, characters, and intention) from information in a text and from world knowledge (Zwaan et al. 1995). Much of this research shows that aspect affects what portions of basic event schemas become available during reading. For example, Mozuraitis et al. (2013) showed that aspect information is encoded moment-to-moment, influences the integration of subsequent events into the discourse model, and is modulated by world knowledge regardless of age. Becker et al. (2013) demonstrated that the three temporal event properties of grammatical aspect, lexical aspect, and pragmatic duration constrain readers’ representations of narratives. Hart & Alberracin (2010) focused on how the intention of agents is more available when an action is conveyed in an imperfective aspect than a perfective aspect. Ferretti, Rohde, Kehler and Cruthley (2009) discovered a goal bias in transfer-of-possession events mediated by perfective aspect. Ferretti, Kutas and McRae (2007) showed that spatial information associated with an event is more available with an imperfective than a perfective aspect. Madden and Zwaan (2003) demonstrated the impact of aspect on readers’ situation models through the use of visual images. Finally, according to Magliano and Schleich (2000), readers interpreted imperfective events as on-going, and perfective actions as completed when
asked about events’ completion status. Aspect also interacted with world knowledge about the typical duration of events.

To be able to determine whether NNS pay attention to aspectual meanings in narratives, it is helpful to understand how college age NS cognitively track such meanings. Native English readers treat morphosyntactic aspect as instructions how to cognitively process a text. In LTM, readers will sequence discourse situations temporally in the case of perfective aspect or treated them as simultaneous when imperfective aspect is used (Becker et al. 2013; Ferretti et al. 2007; Ferretti et al. 2009; Hart & Albarracin 2009; Madden & Zwaan 2003; Magliano & Schleic 2000; Morrow 1985a, 1985b, 1986; Mozuraitis, Chambers & Daneman 2013). For example, Progressive imperfective sentence (2a) She was passing a truck in Table 1 above is thus treated as simultaneous in LTM and interpreted as concurrent with the next sentence, sentence (3) Suddenly she had to swerve to the right. Inferences can be drawn that upon swerving, the protagonist likely will hit the truck. By contrast, if this sentence is in the perfective aspect, expressed by the Simple Past, i.e., (2a) She passed a truck, it is treated as preceding sentence (3), and the two sentences are interpreted as sequential. Readers are not very likely to draw the inference that she hit the truck.

In WM, aspect impacts how information is activated and whether it stays in focus (Becker et al. 2013; Ferretti et al. 2007; Ferretti et al. 2009; Hart & Albarracin 2009; Madden & Zwaan 2003; Magliano & Schleic 2000; Morrow 1985a, 1985b, 1986; Mozuraitis, Chambers, & Daneman 2013). Both imperfective and perfective situations become activated but only imperfective situations stay activated for up to 4 sentences (Magliano & Schleic 2000) and are in focus for subsequent context even longer (4–7 sentences), while perfective situations are completed (Schramm 2001, 1998). This is more apparent in 2-state situations because of the lexicalized second state. For example, sentence (2a) She was passing a truck will stay in focus longer than (2b) She passed a truck and subsequently can be used for further inferencing. Native English-speaking readers will infer that the protagonist is next to the truck in (2a) while they will infer that she is ahead of it in (2b) because of the lexicalized second state of being past the object of the passing. Consequently, readers may infer that the cause for the strong bump in (4) was that Shelley ran into the truck in (2a) when she swerved in (3), but that inference is not as likely in (2b). Thus, meanings from the second sentence in the imperfective (2a) will more likely be reactivated in WM when readers encounter the surprise effect at the end of the story than meanings from perfective sentence (2b). Aspect can be used in conjunction with 2-state situations to direct the reader’s causal inference-making. Thus it also provides a measure for gauging whether non-native readers notice aspectual meanings in a text while reading and whether they cognitively register them through pragmatic retrieval. Since the sentence structure is fairly
unambiguous, it should not require deep computation of structural information, and no differences in syntactic WM-effects in non-native versus native comprehenders are likely, only morpho-semantic ones (Clahsen & Felser 2006).

5. The present study

While past research has explored the acquisition of linguistic means for expressing aspectual meanings by non-native speakers, has focused on the meaning and functions of morphosyntactic-aspect distinctions, and has investigated the cognitive processing of aspectual meanings in native speakers, these efforts have usually not been woven together to study comprehension in a narrative context. For example, to study meaning acquisition, Bardovi-Harlig (2000, 2014) looked into the tense-aspect markers used to express temporality in narratives. Likewise, Klein (1995) investigated evolving temporality in learners using narratives and retell tasks. These studies have provided significant insights into our understanding of the acquisition of aspectual meanings in narratives, but each interlanguage stage affords learners with access to different linguistic means and concurrent processing capabilities connected to aspectual meanings, and we currently do not have a good understanding of whether advanced learners process such meanings, whether they do so without explicit instruction, and how closely their processing resembles native-speaker processing.

To deal with this lack of interlanguage processing data, the current research investigates the comprehension of morphosyntactic-aspect meanings and their narrative use in non-native speakers and compares it to the cognitive processing of such meanings in native speakers all within the same experiment. In the experiment, a target 2-state situation is presented in the imperfective or perfective viewpoint aspect early in a story (see Table 2). After several intervening situations, the target situation is re-invoked by a surprise situation caused by it. The aspect-mediated potential relevance of the target situation is measured by word completions (an online measure to access the activation level of information in memory; e.g., Tiggemann et al. 2004) immediately after the target and directly before and after the surprise effect. The number of word completions reflects how easily a concept can be accessed in memory. Concepts from situations with less semantic and pragmatic relevance produce fewer word completions than concepts with greater relevance. In the present research, memory access to target concepts should be easier (more frequent word completions) to the degree that the concept is already active in the situation model. We expect this to vary depending on the viewpoint aspect of the predicate (imperfective vs. perfective) for both groups of speakers (NNS and NS).
Table 2. Example of the combinations of viewpoint aspect and causal structure

<table>
<thead>
<tr>
<th>Causal Structure</th>
<th>Imperfective</th>
<th>Perfective</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-State Cause</td>
<td>Was passing the truck</td>
<td>Passed the truck</td>
</tr>
<tr>
<td>Surprise Effect</td>
<td>Felt a strong bump</td>
<td>Felt a strong bump</td>
</tr>
</tbody>
</table>

Based on previous NS findings, we expect a main effect of aspect in LTM (but not WM as discussed below) because, in LTM, an imperfective situation is interpreted as ongoing while a perfective situation receives a completed interpretation. Questions at the end of stories explicitly query causal content. They are included to confirm in LTM that aspectual meanings and their use were processed. By contrast, the situation in WM is more complex. A discourse concept is not expected to be more available directly after a cause when its situation of origin is in the imperfective compared to when it is in the perfective viewpoint aspect. This will be measured through word completions immediately before and after the target situation. We also expect that any potential subsequent availability effect of viewpoint aspect will subside after several intervening situations. Specifically, 5–7 sentences after the target situation the activation of the target concept is measured by word completion directly before the surprise situation. We predict little target concept activation even though the imperfective situation information will be stored as in focus and can be accessed even after four intervening situations. Then it is measured again immediately after the surprise situation to determine whether the integration of the surprise effect with the preceding context generates an availability advantage for in-focus concepts from an imperfective target situation with a cause. We thus predict a main effect of location (before vs. after target cause and surprise effect situations) and an interaction between aspect and location. This is because imperfective situation concepts are neither more available before or after target situations nor before surprise situations, only after surprise situations. We also anticipate carrying out planned comparisons to examine the potential interactions allowed for by the factorial design.

Predictions regarding whether non-native speakers process situations presented in these two different viewpoint aspects in narratives differentially are also intricate. The predictions for non-native speaker processing must consider the explicitness with which aspectual meanings and their use are presented. As stated earlier, learning without explicit instructions, or unaware learning, has been considered a possibility. The current study presents aspectual meanings communicatively (without explicit instruction) by embedding them in narrative texts. We expect effects of the two different viewpoint aspects during their processing in WM in this implicit narrative context that parallel those from NS. This means
that we do not expect a discourse concept to be noticed (consciously attended to) and to be more available to word completions immediately after the cause when stemming from an imperfective (rather than a perfective) target situation. Planned comparisons will also be conducted to investigate the availability of concepts in both aspects before and after the target situation; we expect greater availability afterwards regardless of aspect because these concepts were processed immediately prior. Additionally, data that is noticed may or may not become intake (registered cognitively during retrieval from LTM) that can be converted into learning in later stages of the acquisition process. Intake memory of such content will also be studied via word completion in WM, where in the imperfective condition we expect more frequent word completions after the surprise effect re-activates the target situation with the cause from LTM; and via question-answering in LTM, where in the imperfective condition we expect more frequent answers that the target situation was the cause for the surprise. We expect both because situations in the imperfective aspect have to be integrated into the situation model and stored as in focus in order to be registered during retrieval from LTM for word completion and question-answering. Again, we therefore expect the same effects of viewpoint aspect for NNS and NS in WM and LTM. Planned comparisons will be conducted to investigate the different availability of the two aspects in WM before and after the surprise situation and in LTM at the end of narratives.

Finally, given the difficulty in acquiring aspect that learners experience (Andersen & Shirai 1996; Bardovi-Harlig 2000), that acquisition of tense-aspect is difficult to correlate with exact proficiency levels and occurs in stages (e.g., Bardovi-Harlig 2000), and that high-proficiency (not advanced) learners are known to rely on morphological-semantic information during processing (Clahsen & Felser 2006), one could predict some differences between how NS and NNS process aspect. According to the Aspect and Interlanguage Discourse Hypotheses (Bardovi-Harlig 2000), perfective aspect and discourse foregrounding evolve first, followed by imperfective aspect and backgrounding. Also the former is initially associated with 2-state situations like the ones used in the present study, the latter with 1-state situations. As half of the 2-state target situations in our stories are in the perfective aspect, the combination acquired first, whereas the other half is in the dispreferred imperfective, learners might show a processing effect more strongly with perfective situations. However, these predicted differences are based on previous research in language production. Production may lag considerably behind comprehension (Swain 1985, 1995). We therefore predict, as stated above, that a comparison with NS comprehension shows that NNS discourse concepts are also more available in WM after surprise effect situations that follow imperfective target causes, and that there are more frequent answers with information from imperfective target situations in LTM just as with NS.
6. Method

6.1 Participants

There were two groups of participants, group 1 consisting of English NS and group 2 consisting of Saudi Arabian NNS. In group 1, 32 native English-speaking undergraduate students (15 females, 17 males) from a large research university in the Upper Midwest participated for a small stipend or credit in an introductory psychology course. In group 2, there were 25 Arabic-speaking students (four females and 21 males), attending a nation-wide college preparation language program at a small private university in a large metropolitan area in the Upper Midwest. Their proficiency distribution on a 12-level rating scale, as determined by the institution’s own placement test (listening, speaking, grammar, and writing) that is given to students on their first day, was eight students on the high-advanced level (levels 10–12), 15 students on the advanced level (levels 8–9), and two students on the low advanced level (level 7) (M = 9.24, SD = 1.51). Their ages ranged from 19 to 48 (M = 23.44, SD = 6.49). They had been in the U.S. between four and 20 months (M = 11.48, SD = 4.66), and their formal schooling in English ranged from four to 144 months (M = 50.32, SD = 34.91). Their self-assessed casual speaking skills on a scale of 1–10 were M = 6.88, SD = 1.64; academic speaking skills M = 7.36, SD = 1.41; academic listening skills M = 8.12, SD = 1.36; academic reading skills M = 7.60, SD = 1.29; and academic writing skills M = 6.96, SD = 1.70.

6.2 Design

The experiment contained a 2 (speaker: English native and Arabic non-native English) x 2 (aspect: imperfective and perfective) x 2 (causal operator: cause and effect) x 2 (probing location: before and after) mixed design. There were two kinds of English speakers: speakers either spoke English as a first language (i.e., native English) or Arabic as a first language (i.e., non-native English). In addition, there were two aspects. A critical aspect sentence contained a 2-state situation that was either in the imperfective (i.e., Past Progressive) or perfective (i.e., Simple Past) viewpoint aspect. There were four locations at which the word completion prompt

---

5. Arabic, unlike English, does not have morphosyntactic imperfectives (Fassi Fehri 2012). Future studies will have to explore the effects of the current experimental manipulations on learners of English from first languages with morphosyntactically marked imperfectives, such as French or Spanish.

6. Even though there is a wide range in age and schooling, we are confident that the measure most important here, i.e., proficiency, was based on a robust assessment process.
could appear. Specifically, it could appear either immediately before or immediately after both the cause and the final-effect sentences. There were two kinds of causal operators depending on whether a mid-narrative cause or an end-of-narrative surprise effect was expressed. Speaker and causal operator were between-subject variables, and aspect and probing location were within-subject variables. Additionally, there was a comprehension question which appeared after a prompt indicating the end of a story.

6.3 Passages

16 narratives were designed to test the effect of grammatical aspect on memory in NNS and NS, which has been investigated before in various contexts (Becker et al. 2013; Ferretti et al. 2009; Ferretti et al. 2007; Hart & Albarracín 2009; Madden & Zwaan 2003; Magliano & Schleich 2000; Morrow 1985a, 1985b, 1986, 1990; Mozuraitis et al. 2013). In addition, situation aspect was held constant to include only sentences with telic 2-state (with endpoint) situations. The following narrative-structure template was used: a pair of cause situations (Cause 1 and 2) and a single surprise-effect situation (cf. Table 3 below). The two causes are such that a causal link may be inferred between either cause and the effect. When Cause 1 is in the imperfective aspect, we expect both Cause 1 and Cause 2 to be potentially inferred as causes for the surprise effect. We predict that using imperfective aspect will make Cause 1 more accessible in both WM and LTM because only part of its

Table 3. Sample Narrative “Shelley and the Deer”

<table>
<thead>
<tr>
<th>Narrative text</th>
<th>Causal Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Shelley was visiting her grandparents in Northern Minnesota.</td>
<td></td>
</tr>
<tr>
<td>2) It was a beautiful drive,</td>
<td></td>
</tr>
<tr>
<td>3) the narrow two-lane road lined with trees,</td>
<td></td>
</tr>
<tr>
<td>4) leaves beginning to turn.</td>
<td></td>
</tr>
<tr>
<td>5) The road was empty except for this slow poke in front of her.</td>
<td></td>
</tr>
<tr>
<td>6) Impatiently, she <em>passed/was passing</em> the pickup.</td>
<td>Cause 1</td>
</tr>
<tr>
<td>7) Unexpectedly, a deer came charging out of the woods</td>
<td>Cause 2</td>
</tr>
<tr>
<td>8) forcing her to yank the steering wheel to the right.</td>
<td>Cause 2</td>
</tr>
<tr>
<td>9) Within split seconds, several thoughts raced through her head.</td>
<td>Filler</td>
</tr>
<tr>
<td>10) The car belonged to the company her dad works for.</td>
<td>Filler</td>
</tr>
<tr>
<td>11) Would she get in trouble for taking it without asking?</td>
<td>Filler</td>
</tr>
<tr>
<td>12) Would he punish her by not letting her use his car again?</td>
<td>Filler</td>
</tr>
<tr>
<td>13) She felt a strong bump.</td>
<td>Effect</td>
</tr>
</tbody>
</table>
pre-state without an end point is expressed, and it additionally has a lexicalized post-state, thus keeping it in focus.

When Cause 1 is in the perfective aspect, we expect that Cause 2 is more likely to be inferred as the cause for the Effect. Cause 2 meets the causal criteria, functions as the “fall-back” cause, and avoids that readers have to infer the first cause regardless of aspect for lack of an alternative. A perfective Cause 1 does not meet the causal criteria as readily because it is not in focus anymore as a result of its aspect. It is not very likely to be inferred as a cause.7

Table 3 above illustrates the two inferencing scenarios in the narrative *Shelley and the Deer*. Clause 6 contains the first cause with the aspect manipulation, shown in italics. Clauses 7 and 8 contain Cause 2. Clauses 9–12 constitute fillers between causes and effect to make sure Cause 1 is no longer in WM. Clause 13 constitutes the Surprise Effect and is preceded by a coherence break.

16 texts with a word (stem)8 completion task (e.g., Tiggemann et al. 2004) were used to test moment-to-moment processing. This measure investigates whether readers attend to, or “notice”, aspectual meanings, and thus causal inferencing impacts WM. The word completion task, during which readers complete partial words (stems) on a separate page, provides access to the level of activation of information in memory. Subjects are presented with texts containing the words in question and, at fixed points in the narrative, are then asked to complete such two-letter stems that contain the initial letters of the words presented before. The number of stems completed with words from the two text versions is the measure of the word completion task. The partial word was part of the verb phrase, either a direct or prepositional object used in the sentence that expressed the Cause 1 situation. In the sample text in Table 3 the partial word was “P I _ _ _ _” for *pickup*. Each partial word from the story was also complemented with another partial-word distracter to avoid having participants develop response strategies. Participants were asked to complete the words as quickly and completely as possible. Partial words were presented at exactly one of four potential locations in each text: immediately before or immediately after the two causal operators Cause 1 and Effect. Thus there was only one relevant partial word per story. Presenting partial words directly before and after the two causal operators also provides a measure to gauge whether participants are comprehending the texts regardless of aspect.

---

7. A causal relation between the effect and the poststate of a perfective Cause 1 is possible. But this causal relation is not nearly as likely to be inferred because the causal connection is very weak.

8. We put “stem” in parentheses because we recognize that “word stem” is the customary term for this experimental task. While we acknowledge this use, we do not want to imply that only morphological stems in the linguistic sense were used as partial words for this task.
For the moment-to-moment part of the experiment, the independent variables linguistic aspect, i.e., perfective or imperfective situation, and location of word completion, i.e., right before or after Cause 1 or Effect, were randomized. Each text contained only one instance of each variable and half the participants saw texts with a probe at Cause 1, half with it at the Effect. The dependent variable was number of completed partial words. Word completions were coded according to number of partial-word completions. Every time participants completed a partial word from Cause 1 correctly, a score of 1 was assigned, indicating that the situation meaning from Cause 1 was activated in memory. If a partial word was not completed correctly or at all, 0 was assigned. For the offline portion, each text contained one comprehension question that participants needed to answer. Answers were coded according to number of mentions of Cause 1 meanings or of non-completion of Cause 1 situations. If participants referred to information from the first cause with the aspectual meanings or to its non-completion, a score of 1 was assigned, indicating that Cause 1 information was accessible in LTM. For example, a score of 1 was recorded if participants mentioned the pickup or indicated that Shelley’s passing of the vehicle was not finished. Similarly, if participants did not refer to information from the first cause or indicated that it had been completed, a score of 0 was assigned. For example, a score of 0 was recorded if there was no mention of the pickup or participants indicated that Shelley had already passed it.

6.4 Procedure

The texts were presented in small color-coded paper booklets. Reading was self-paced, and participants were instructed to read as normally as possible where “normal” was described as how one generally reads fiction. Participants were instructed not to turn back any of the pages to review previous text. After 5–7 sentences in a story, participants in the Before-Cause condition flipped the page and encountered a partial word; in the After-Cause condition, the partial word followed immediately after Cause1. In the Before-Effect condition, participants encountered the partial word another 5–7 sentences after Cause1 and right after the Surprise Effect in the After-Effect condition. Participants were instructed to complete the word task as quickly as possible and in no more than 15 seconds. At the end of the story, participants were presented with the asterisks and then flipped the page to the question. Instructions were to answer the question as best as possible, and if participants did not know the answers after 30–45 seconds to put down a dash and move on to the next story.
7. Results

7.1 Moment-to-moment processing

The dependent variable was the probability of completing the partial word from the Cause 1 situation in which aspect was manipulated. Mean probabilities were calculated for both NNS and NS at each of the four locations: before and after Cause 1 and before and after the Effect. The results for the imperfective (Past Progressive) and perfective (Simple Past) versions for both groups of participants are reported in Table 4.

<table>
<thead>
<tr>
<th>Location</th>
<th>Imperfective</th>
<th>Perfective</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Non-native Speakers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Cause</td>
<td>.12</td>
<td>.13</td>
</tr>
<tr>
<td>Post-Cause</td>
<td>.40</td>
<td>.28</td>
</tr>
<tr>
<td>Pre-Effect</td>
<td>.38</td>
<td>.29</td>
</tr>
<tr>
<td>Post-Effect</td>
<td>.46</td>
<td>.37</td>
</tr>
<tr>
<td>Native Speakers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Cause</td>
<td>.30</td>
<td>.23</td>
</tr>
<tr>
<td>Post-Cause</td>
<td>.69</td>
<td>.34</td>
</tr>
<tr>
<td>Pre-Effect</td>
<td>.41</td>
<td>.22</td>
</tr>
<tr>
<td>Post-Effect</td>
<td>.61</td>
<td>.32</td>
</tr>
</tbody>
</table>

A 2 (speaker) x 2 (aspect) x 2 (probing location) x 2 (causal operator) mixed ANOVA was conducted on the completion frequencies of partial words. There was a significant main effect for speaker, $F(1, 53) = 6.31, MSE = .144, p = .01, \eta_p^2 = .11$, and location, $F(1, 53) = 16.40, MSE = .110, p < .001, \eta_p^2 = .24$. Words had a higher probability of being completed by NS than by NNS and after a causal operator (cause or effect) than before. Aspect approached significance, $F(1, 53) = 2.80, MSE = .144, p = .10, \eta_p^2 = .05$. The interaction between aspect and location was marginally significant, $F(1, 53) = 3.76, MSE = .06, p = .06, \eta_p^2 = .07$. Planned comparisons were conducted to assess the nature of this interaction. There was, most importantly for this study, a highly significant difference between aspects in the after-operator location, $t(56) = 2.72, p < .005$, but not before.
Because of the unexpected difference in nativeness of speakers, the NS data was investigated separately as a baseline for comparison with NNS data. A 2 (aspect) x 2 (probing location) x 2 (causal operator) mixed ANOVA was conducted on the completion frequencies. There was a significant main effect for location, $F(1, 30) = 12.68$, $MSE = .134$, $p = .001$, $\eta_p^2 = .30$. Words had a higher probability of being completed after causal operators than before. The difference in word completions after imperfective and perfective aspect was marginally significant, $F(1, 30) = 3.66$, $MSE = .07$, $p = .06$, $\eta_p^2 = .11$. Words after imperfective situations were more likely to be completed than after perfective ones. The interaction between aspect and location approached significance, $F(1, 30) = 3.00$, $MSE = .05$, $p = .09$, $\eta_p^2 = .09$. Planned comparisons were conducted to assess the nature of this interaction. Before cause and effect combined or individually, there was no significant difference in the frequency with which native readers completed partial words from situations with different aspects. This difference reached significance in the after-location, $t(31) = 2.74$, $p = .005$. Specifically, the difference approached significance after the cause, $t(31) = 1.41$, $p = .09$ and reached significance after the effect, $t(31) = 2.54$, $p = .01$, with completion values for the imperfective higher than the perfective in both cases. A 2 (aspect) x 2 (probing location) x 2 (causal operator) mixed ANOVA of the NNS data only showed a main effect of location, $F(1, 23) = 5.22$, $MSE = .08$, $p = .03$, $\eta_p^2 = .19$, with words from situations more likely completed after causes and effects than before. There was no significant main effect of aspect.

7.2 Off-line processing

The dependent variable was the probability of answering the causal inferencing question such that the aspect situation (i.e., Cause 1) was considered the cause for the unexpected effect. Mean probability scores for each of the four conditions were calculated for both NNS and NS. The results for the ongoing imperfective and completed perfective versions for both groups of participants are reported in Table 5.

Table 5. Mean Frequencies and Standard Deviations of Cause 1 Mention in Questionnaire Answers by Non-Native and Native Speakers

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Aspect</th>
<th>Imperfective</th>
<th></th>
<th>Perfective</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>Non-native Speakers</td>
<td></td>
<td>.17</td>
<td>.19</td>
<td>.18</td>
</tr>
<tr>
<td>Native Speakers</td>
<td></td>
<td>.45</td>
<td>.29</td>
<td>.34</td>
</tr>
</tbody>
</table>
A 2 (speaker) x 2 (aspect) x 2 (probing location) x 2 (causal operator) mixed ANOVA was conducted on the answer frequencies. There was a significant main effect for speaker, $F(1, 53) = 48.63, MSE = .06, p < .001, \eta^2_p = .48$. Situations had a higher probability of being mentioned by NS than by NNS. Aspect did not reach significance ($F < 1$). Most important for this study, the aspect-by-speaker interaction was marginally significant, $F(1, 53) = 3.29, MSE = .05, p = .07, \eta^2_p = .06$. The aspect-by-location interaction approached significance, $F(1, 53) = 3.01, MSE = .05, p = .08, \eta^2_p = .05$, and the 4-way interaction between speaker, aspect, probing location, and causal operator was marginally significant, $F(1, 53) = 3.48, MSE = .05, p = .06, \eta^2_p = .06$. Planned comparisons were conducted to assess the nature of the two-way interactions. As for the aspect-by-speaker interaction, there was virtually no difference in NNS in frequency of answers from the two different aspectual conditions ($F < 2$). By comparison, there was statistically significant higher frequency of imperfective than perfective-cause answers in NS, $t(63) = 2.45, p < .01$. This unexpected interaction was further explored with planned comparisons of NNS and NS. For NS, there was a main effect of aspect, $F(1, 30) = 5.18, MSE = .07, p = .03, \eta^2_p = .15$, while that was not the case for NNS.

As for the aspect-by-location interaction, the aspectual difference was higher in the before-locations for all speakers combined, $t(56) = 2.54, p < .01$. These unexpected results from both the two-way and four-way interactions are discussed below.

8. Discussion

8.1 Moment-to-moment processing

Results of the moment-to-moment part of the experiment indicate that both NNS and NS process the narrative situations with the aspectual manipulation; information from such situations with causes was more accessible after reading them and later situations containing surprise effects than before either. Importantly, with regard to the purpose of this experiment, only NS readers draw on aspect in understanding the temporal and thus causal structure of narratives. This finding is consistent with both the claim that there may be differences in fluency and automaticity in processing morphological-semantic information between advanced,
yet not highly proficient, NNS and NS (Clahsen & Felser 2006) because aspect is difficult to acquire for language learners (Andersen & Shirai 1996; Bardovi-Harlig 2000) and that aspect and its morphosyntactic carriers convey cognitive-processing instructions as to how information in working memory is to be managed (Becker et al. 2013; Carreiras, Carriedo, Alonso & Fernandez 1997; Ferretti et al. 2009; Ferretti et al. 2007; Garnham et al. 1995; Givón 1995; Hart & Albarracin 2009; Kintsch 1995; Madden & Zwaan 2003; Magliano & Schleich 2000; Morrow 1985a, 1985b, 1986, 1990; Mozuraitis et al. 2013; Sanford & Garrod 1981). This data is also consistent with studies suggesting that aspectual meanings are initially activated (Becker et al. 2013; Mozuraitis et al. 2013; Magliano & Schleich 2000) and that this activation decays after 4–6 sentences. Furthermore, it also indicates that decayed meanings tagged as in-focus can still be reactivated in WM after one intervening sentence (Mozuraitis et al. 2013) and, in fact, after up to 4–6 intervening sentences (Schramm 1998). It differs from Magliano & Schleich (2000) and Schramm (1998, 2001) and confirms Mozuraitis et al. (2013) in that the sensitivity to aspect appeared to be immediate. This may be a function of the combination of narrative context and task. Narrative context disfavors the use of imperfective aspect as part of the narrative plot (Hopper 1982), which may have resulted in additional activation of imperfective meanings even though in principle establishing the temporal relationship between adjacent narrative events is semantically equivalent in the two aspects; and it may be a function of task (Rapp & Mensink 2011), as Magliano & Schleich (2000) and Schramm (1998) used question-answering and speeded-recognition tasks, while Mozuraitis et al. (2013) employed an eye-tracking paradigm, each of which constitutes a different context potentially with different processing strategies.

As a separate group, however, the advanced NNS struggled with aspectual meanings as observed before (Bardovi-Harlig 2000; Andersen & Shirai 1996). One of the proposed second-language acquisition paths, that of unaware or low-level awareness (Long 1991; Tomlin & Villa 1994; VanPatten 1990), should allow NNS to achieve behavior resembling that of natives. Consistent with the unawareness hypothesis, readers should exhibit an activation profile that puts imperfective situations in focus and accesses them during causal inferencing without receiving neither any instruction nor other feedback or input besides the narrative itself. Post-effect activation is not consistent with the unawareness path. NNS participants were not confident that imperfective situations were in focus. However, they did activate meanings after the cause and effect regardless of aspect. This is consistent with the view that they were generally considering the semantic content of the situations in question, but not aspectual meanings. In contrast, for readers with NS fluency and automaticity, processing aspectual meanings in narratives is a common activity and provides a baseline of activation patterns for how information is activated in WM.
8.2 Off-line processing

Off-line results clearly indicate that NNS did not exhibit the familiar aspectual pattern, and NS did, where situations in the imperfective aspect are more available in LTM, i.e., they are more in focus in the narrative (Becker et al. 2013; Carreiras et al. 1997; Hart & Albarracin 2009; Ferretti et al. 2009; Ferretti et al. 2007; Magliano & Schleich 2000; Mozuraitis et al. 2013; Morrow 1985a, 1985b, 1986, 1990; Schramm 1998, 2001). More specifically, NNS did not appear to utilize the semantic information from either aspect during the construction of temporal and contingent causal relationships in the situation models of narratives. Unexpectedly, the difference in likelihood that situations were considered the cause of a later event increased for the combined group of readers when probing occurred before both causal operators (cause and surprise effect). This suggests that in the before-location the combined group had a higher likelihood of updating their situation models (mental representations of text events) to specifically indicate that a situation was imperfective since the aspectual difference was significant in the before location. However, the difference in probability scores between aspectual conditions was flat for NNS across all probing locations and text versions. These results are somewhat surprising. Perhaps NNS readers were generally (regardless of the aspect used) less confident about the aspectual status of situations after reading causal-operator sentences, and the aspectual difference in the before-location was due to NS behavior. To assess the extent to which speakers’ nativeness status played a role, planned t-tests were conducted. The likelihood scores that the aspectual difference was higher in the before-locations were indeed less than chance only for NS, $t(31) = 2.38, p = .01$. Thus, NNS participants were inferencing similarly about the aspectual status of all situations regardless of probing location. Only NS drew off-line inferences based on aspect in the before probing location, presumably because of the specific experimental task (Rapp & Mensink 2011).

Regarding the speaker-aspect-location-operator interaction, it appears that in addition to speaker and location, operator (cause and effect) also influenced the likelihood of readers drawing causal inferences based on aspect. Perhaps a situation exists that is similar to the interaction between aspect and location. To assess whether the native status of speakers influenced inferencing at the cause versus surprise effect, a series of t-tests were conducted. A trend existed in NNS for mean values to be lower for cause than for effect conditions regardless of aspect. NNS participants’ likelihood of naming the cause in answer to questions increased regardless of aspect after seeing a probe at the surprise effect. Readers’ situation model construction may have been affected by disruption through probing early in narratives. The impact of probing was less when the interruption occurs right before the end of the story presumably because the cause has entered the situation model already.
NNS answers to queries about their situation model content regarding imperfective causes are inconsistent with the unawareness hypothesis (Long 1991 & 1998; Tomlin & Villa 1994; VanPatten 1990). According to the unaware path to language acquisition, non-native readers’ situation model should exhibit evidence of aspectual meanings and causal inferencing based on the contextualized presentation of the linguistic features. It was not evident from answers that imperfective situations were ongoing and perfective situations closed off.

9. General discussion and conclusion

The results of this study indicated that some awareness in the form of attention to aspectual meanings in working memory appears to be required for possible second-language acquisition. This is consistent with prior research on grammatical forms by Hama & Leow (2010), Leow (1997), Rosa & Leow (2004), Rosa & O’Neill (1999), Sachs & Suh (2007), and Slabakova (2014). In this study we demonstrated that advanced non-native English speakers did not notice (i.e., consciously attend to) aspectual distinctions. It has been claimed that, during language acquisition, readers may or may not attend to incoming information from a text by default. Learning may occur implicitly through language presented in its natural context or explicitly when language presentation is altered from its normal communicative context. When the language feature to be learned is aspect and the context in which it occurs is narrative, learners must create a causally coherent representation. Situations that are presented in the perfective aspect are strung together sequentially with each situation typically causing the next to occur. Imperfective situations are fairly likely to suspend this process and be considered as remaining in focus and available for causal incorporation later.

Results indicated that learners cannot make this aspectual distinction in a naturally occurring narrative as the communicative context. Situations conveyed with either aspect appear to have the same impact on activation in working memory. The current narrative context does not result in attention to aspect. And readers have an equal likelihood of updating their situation models to indicate that imperfective and perfective situations have been considered. Results also indicate that situation model construction based on story situations with causal information appears to occur fairly locally. Advanced learners may have limited resources for story situations at a level of activation needed to establish narrative coherence. It appears activation decays while working-memory data is collected, and situation models are updated to indicate that situations are less relevant. This is consistent with claims about low-span readers’ (Just & Carpenter 1992) and language learners’ (Ortega 2009; Cowan 2005) limited memory capacity.
Learners may not have adequate working memory resources needed to attend to the experimental task.

Low-level awareness, known as intake (cognitive registration during retrieval from LTM (Leow 2000; VanPatten 1993, 1997; Williams 2004)), also appears to be required for the construction of situation models during acquisition of aspect. This is consistent with prior research on the acquisition of grammatical forms (Leow 1997; Hama & Leow 2010; Rosa & Leow 2004; Rosa & O’Neill 1999; Sachs & Suh 2007; Slabakova 2014). The low likelihood of retrieval of either cause (compared to native English speakers) suggests that unaware learning of aspect in a narrative context, if there was any, does not impact the updating of learners’ situation model. As they are building the causal text representation, readers do not appear to pay attention to the aspectual difference when retrieving information from the situation model, or they may not have stored it at all. In either case, this lack of retrieval suggests that intake and subsequent acquisition of aspect comprehension does not occur at this stage of the acquisition process.

The results of this study contribute to the research on the acquisition of meaning. Both the approach of investigating linguistic forms by themselves (focus on forms) or in meaningful contexts (focus on form) (Hama & Leow 2010; Slabakova 2014; Williams 2005) and the concept-oriented or meaning-oriented approach (Bardovi-Harlig 2014, 2000; Klein 1995) provide descriptions on how the acquisition of meaning proceeds during language production. Our results indicate that learners considered meanings also during comprehension, not only in production. The activation level of such situations was at a higher level right after reading sentences with them (regardless of aspect). This is consistent with the view that situation meanings had an impact on reader’s attention during reading. Similarly, learners were sensitive to narrative meanings immediately after they finished reading. However, their likelihood of updating their situation model after causes early in stories to indicate later that situations were read is reduced through interference, presumably because of inadequate WM resources. This limitation has been noted before (Cowan 2005; Just & Carpenter 1992; Ortega 2009) and appears to apply during the acquisition of meaning as well.

It has been argued that very high-proficiency learners rely on morphological-semantic information during processing in the same way as native speakers, but less so on syntactic cues (Clahsen & Felser 2006). Furthermore, we know that learners acquiring a semantic concept such as aspect move through stages, pragmatic first, then lexical, finally morphological, at which they utilize different means of expression (Bardovi-Harlig 2014). Aspect provides an important diagnostic tool for the comparison between native and non-native speakers at different acquisition stages of the means of expression for aspect. More specifically, lexical aspect provides insights into the interplay of semantic and lexico-syntactic cues at the
sentence level (Klein 1994; Ryle 1949; Schopf 1984; Vendler 1957). And viewpoint aspect is comprised of information from predicate morpho-syntax and semantic meanings (Comrie 1976; Klein 1994; Schopf 1984; Smith 1997). In terms of this study, advanced non-native readers do not appear to possess the adequate linguistic resources needed to understand the complex interplay between situation (lexical) and viewpoint (grammatical) aspect to maintain imperfective narrative situations in focus and subsequently make relevant causal inferences. Learners may not be at a high enough proficiency level to have access to the required morphological-semantic information, let alone the syntactic resources, compared to the ones in Clahsen and Felser (2006). However, narratives in this study are not very complex syntactically compared to the syntax in Clahsen and Felser (2006). Instead, it is more likely that the interplay between narrative lexico-syntax and morpho-syntax and their meanings is not available to learners at this stage as a resource. In fact, a recent study indicated that learners at this stage took note of verbal morphosyntactic aspect information when it was visually highlighted in texts, but semantic and pragmatic understanding nevertheless did not ensue (Meidal 2008). Learners in the Meidal study and in the current study appear to be between the lexical and morphological stages of aspect acquisition (Bardovi-Harlig 2014). Because of the compositional nature of the means of expression for aspect, learners can neither solely rely on the order of events nor on adverbials or connectives when reading imperfective situations needing to be kept in focus and available for subsequent causal inferencing during narrative comprehension and construction of the situation model. The exact nature of resources available at this stage of acquisition requires further investigation, in part to ascertain appropriate instructional strategies.

The results of a recent preliminary study suggest some possibilities for instructional strategies for aspect (Kivimagi 2013). First, explicit semantic and pragmatic information about aspect was provided and readers were encouraged to pay attention to visually enhanced words in the narratives. Explicit instruction included a metalinguistic explanation of both aspects (completed nature of perfective and ongoing nature of imperfective aspect), their impact on discourse (bounded nature of perfective and unbounded nature of imperfective aspect situations), and how to make causal inferences. The didactic steps consisted of instructor-modeling, guided practice, and independent practice. After less than two hours of instruction, learners’ likelihood of having stored imperfective situations as in-focus was significantly higher both in working memory and in the situation model when compared to perfective situations. Instruction of semantic and pragmatic aspectual meanings appears to have had some impact, suggesting that learners’ struggles with resources involved semantic as well as pragmatic, rather than syntactic, information. This seems reflective of the morphological stage of concept acquisition (Bardovi-Harlig 2014) and in line with the findings from this study.
This study also appears in line with the well-established impact of aspectual meanings on memory activation and situation model construction of native English speakers (Becker et al. 2013; Carreiras et al. 1997; Ferretti et al. 2009; Ferretti et al. 2007; Hart & Albarracin 2009; Madden & Zwaan 2003; Magliano & Schleich 2000; Morrow 1985a, 1985b, 1986; Mozuraitis et al. 2013). In this study we demonstrated that one of the functions of aspect appears to be to convey temporal information regarding the establishment of causal links between narrative situations. Readers follow the norm that the order of situations mirrors their chronological order and that unless indicated otherwise events are completed and successive on a timeline (i.e., they follow the iconicity assumption (Dahl 1984) or the principle of natural order (Klein 1987)). Aspect indicates whether this norm should or should not be followed. Situations presented in the imperfective aspect had a much higher likelihood to be considered temporally available and therefore causes of subsequent surprise effects than situations in the perfective aspect, both during processing and in the resulting situation model. However, readers cannot maintain imperfective situations as available in WM indefinitely. Instead, situation aspect information from the sentence with the relevant causal situation seems to be stored in-focus in the situation model. Readers subsequently appear to establish causal links with in-focus situations, as information from the situation with the effect is introduced, and ongoing situations are presumably stored as completed. NS data indicates that viewpoint aspect affects the availability of situations for causal inferencing during narrative reading.

Acknowledgements

The authors would like to thank Martha Bigelow, the Text Group at the University of Minnesota, and several anonymous reviewers for their helpful suggestions on this article, C. Randy Fletcher for help with the design of and participant recruitment for the experiments, and Margaret Pederson, Mary McKay, and Catherine Mason at ELS of St. Thomas University for access to their non-native English students. However, any shortcomings in this article are our own responsibility.

References


Processing of aspectual meanings


This study shows that using computational linguistic models is beneficial for descriptive linguistics and psycholinguistics. It applies two models to various English genres and learner language: 1) surprisal and 2) a syntactic parser, allowing us to investigate the role of ambiguity and the interplay between idiom and syntax principles. We find that surprisal and ambiguity are higher for learner language, while parser scores and model fit are lower. In addition, the random application of alternations leads to more ambiguous sentences. Failures to generate optimal orderings in the sense of relevance theory, such as nonnative-like utterances by language learners exhibit, increase processing load, both for human and automatic processors. As human and automatic parsing difficulties correlate, we suggest syntactic parsers as psycholinguistic processing models.

Keywords: language processing, statistical models, idiom and syntax principle, ambiguity, syntactic parsing

1. Introduction

Linguistic differences in language use according to social factors, genre and register, or historical language change are often reported as differences in relative frequencies of the phenomena under investigation, and significance tests such as chi-square or the T-test are applied. It has emerged, however, that such approaches are unreliable: first, because the factors are not independent, and, second, because the assumptions that the significance tests make are often violated. As a result, the use of statistical models instead of significance tests has been advocated for instance by Evert (2006) or Gries (2006). The current paper demonstrates the benefit of using computational linguistic methods in descriptive linguistics and psycholinguistics. It shows the usefulness of predictive multi-factorial models given an appropriate independent variable reflecting the envelope of variation and presents a
more involved parser model for language processing, with particular emphasis on the role of ambiguity.

While multi-factorial models allow one to take the dependencies between the various modeled factors into account, a problem still remains: as we can never include, or even know, all factors involved, it is important to spend considerable time on the decision which factors and outcomes to model in the first place. Statistical models predict outcomes (dependent variable) given certain factors (predictors, independent variables). We examine two language models which offer a global perspective on language use rather than focusing on single, local choice points, such as alternations, and offer an additional tool to the linguist for describing e.g. different genres and learner levels. The two language models are surprisal (Levy & Jaeger 2007) and a syntactic parser (Schneider 2008).

The main goal of our paper is to show that these statistical models are global models of language processing. They model, on the one hand, the unexpectedness of the continuation of the discourse, and, on the other hand, the amount of ambiguity which needs to be resolved for analysis and interpretation. Both models are inherently gradient and take a wide range of phenomena into account. Surprisal is a suitable model to demonstrate Sinclair’s idiom principle, while a syntactic parser also takes the interaction with Sinclair’s open-choice principle into account. Concerning language acquisition, for first language acquisition it has been shown that lexical-specific idiom-based language use precedes creativity (Tomasello 2000), but for second language acquisition the situation is less clear (Granger 2009; Ellis 2012). We believe that the two models also have psycholinguistic applications and test this on data from language users whose output differs from that of native speakers in interesting ways: language learners (L2). In particular, as will be shown, corrected L2 data has a higher model fit and lower surprisal than original L2 data.

The paper is structured as follows. In Section 2.1 we argue for the use of multi-factorial language models and discuss how the envelope of variation may be of use. In Section 2.2 we introduce the computational linguistics models we apply: surprisal and the syntactic parser. Section 2.3 introduces learner language and the learner corpora used are introduced in Section 3.1. The methods we employ are described in Section 3.2. In Section 4 we present our results showing that learner language, with higher surprisal and lower model fit, is less optimal in terms of information theory, confirming our, as well as Pawley and Syder (1983)’s, hypothesis that native speakers know best how to balance fixedness vs. expressiveness. In Section 5, we use language models to shed light on the question of ambiguity in language: do we form sentences in such a way as to avoid ambiguity?
2. Background and motivation: Language models

The use of language models for linguistics has developed largely independently in two strands: First, based on the recognition that significance testing is not enough for descriptive linguistics, as we discuss in Section 2.1. Second, in computational linguistics, the application-oriented processing of natural language for varied tasks (including speech recognition, part-of-speech tagging, syntactic parsing, or machine translation) has led to the development of a rich array of language models, some of which will be outlined in Section 2.2. All models presented here are statistical models. As such, they are also empirical models, have a vital statistical component, and most of them also have a vital rational component, be it a set of rules (such as unknown capitalized words defaulting to proper name in English part-of-speech taggers) or axioms (such as a set of grammatical relations or the assumption that words are separated by space characters).

2.1 A case for statistical language models in linguistics

Statistical models are very widely used in science, for example in meteorology, architecture, engineering, or physics. Increasingly, the advantages of multifactorial, predictive models have also been recognized in Descriptive Linguistics, e.g. regression models (Evert 2006; Gries 2010; Gries 2012; Gries to appear) or equivalent ANOVA models. Nevertheless, descriptive linguistic approaches often still use significance testing rather than statistical models to describe linguistic differences. In Section 2.1.1 we first discuss why traditional significance tests are inadequate and point to the important factor of genre or subgenre; then, in Section 2.1.2, we consider the advantages and shortcomings of the concept of envelope of variation: on the one hand, it models the choice of language users; on the other hand, often no real choice is actually available.

2.1.1 Significance tests are not enough

Significance tests are important to assess if an observed difference between two varieties, genres or speaker groups is large enough to be statistically significant and not just due to random variation. However, they face a number of serious problems since some of the assumptions they make are often not met.

1. Assumption of random distribution
2. Assumption of independence from other factors
3. Assumption of free choice

2.1.1.1 Assumption of random distribution. The T-test makes the assumption that data is distributed normally. If this were the case, the distance of one occurrence
of a given word to its next occurrence would be normally distributed, as would be its frequency in each document of equal length. However, Church (2000) observes that the probability for a given word to occur is much higher if it has already appeared in the same document before: the assumption of random distribution is jeopardized. Furthermore, Evert (2006) shows that non-random distribution of words does not only apply to content words, but also to grammatical features such as passive forms. In addition, Schneider and Hundt (2012) find that past perfect forms and modal verbs are also not randomly distributed. The histogram of Figure 1 shows that the distribution of past perfect forms per document (of 2,000 words each) in a large subset of the International Corpus of English for Great Britain (ICE-GB) is very different from the Gaussian bell-shaped distribution known as the normal distribution. Accordingly, results of the T-test applied to past perfect forms are generally unreliable.

![Figure 1. Histogram of past perfect form frequencies in ICE-GB](image)

2.1.1.2 Assumption of independence from other factors. While the parametric T-test assumes a normal distribution, the non-parametric χ²-Test does not make such an assumption. However, the χ²-Test does assume that the data are independent of other factors. In Figure 1, while the document itself is a strong factor, individual documents are also influenced by various factors themselves: both by traditional sociolinguistic factors such as age, gender, or social background of the
author, and by the topic and the genre of the document. Hence, the $\chi^2$-Test is equally unsuitable.

Variationist linguistics in particular faces the problem that variation by region is typically much weaker than variation by genre. For example, consider instances of NP postmodification in selected regional components of the International Corpus of English (ICE). Figure 2 presents relative frequencies by genre and Figure 3 shows relative frequencies by region. The differences by region are considerably smaller and this entails that, unless genre is controlled very carefully, regional differences run the risk of being overshadowed by genre differences.

![Figure 2. Relative frequency noun-modifying PPs across genres](image1)

![Figure 3. Relative frequency noun-modifying PPs across varieties](image2)
Moreover, Gries (2010) and Gries (to appear) discuss examples where the application of monofactorial analysis and the application of multifactorial analysis without considering all interactions can lead to incorrect results. He concludes that “multifactorial data must be analyzed multifactorially: [...] the complexities of linguistic data do not reveal themselves easily either to the naked or to the monofactorial eye” (Gries 2010: 143).

2.1.1.3 *Assumption of free choice.* The third problem is that a speaker often has no real choice. The assumption of free choice (as we have named it here) is strictly speaking not a statistical assumption, but a consequence of the influence of factors such as topic and semantics of the document, and in turn often leads to the problem of non-randomness of distributions. The absence of choice is most obvious for content words (Church 2000), where the lexical range is largely predetermined by the semantics of the topic. But also when it comes to choosing function words and other morphosyntactic factors (e.g. tense, voice, aspect and modality), discourse and content often place restrictions. In a narrative, the simple past will be more frequent, while in a scientific paper the resultative present perfect will be used often, and in weather forecasts we expect to hear future tense. In a discussion on normative ethics, modal verbs are inherently very frequent. At best, the speaker in such a discussion has a choice between several modal verbs of obligation (*must, should, need*). Thus, one should approximate speaker choice as envelope of variation (Labov 1969; Sankoff 1988), as we discuss in the following.

2.1.2 *The envelope of variation*

Labov states in what he calls *principle of accountability* that any variable form (a member of a set of alternative ways of ‘saying the same thing’) should be reported with the proportion of cases in which the form did occur in the relevant environment, compared to the total number of cases in which it might have occurred. (Labov 1969: 738)

The question of which envelope of variation to use is also known as the problem of the unit of measurement. If we count frequencies, for example for lexical choice, we model speaker choice as the choice of word *w* (which is under investigation) versus any other word. A more realistic assumption is to count only those contexts in which there is an actual choice for *w* or for one of its synonyms or semantic alternatives on the paradigmatic axis (e.g. antonyms, meronyms).

As mentioned above, Evert (2006) shows that non-random distribution of words does not only apply to content words, but also to grammatical features, passive forms in his investigation. To illustrate this the distributions of raw passive
forms across documents in the Brown Family4 (LOB, FLOB, Brown, Frown) is given in Figure 4.

![Percentage of passive forms per article in Brown family 4](image)

Figure 4. Percentage of passive forms per article in Brown family 4

However, Hundt et al. (2016) discover that a normal distribution emerges if (1) a reasonable envelope of variation is used, and (2) the data are split by the relevant factors. First, instead of using frequencies of passive forms, a better approximation to the envelope of variation is to use each verb as a unit of measurement: a verb is either in the active or the passive form. If one also considers that for many verbs there is no choice, e.g. *be* is never and *have* hardly ever in the passive, and only transitive verb tokens are available as alternation candidates, one can further constrain the envelope. This leads to a reasonable operationalization for the envelope of variation, but, as Figure 5 shows, the data are still not normally distributed – other important factors must be influencing the data.

Second, concerning the influence of other factors, at least the influence of the crucial factor *genre* needs to be taken into consideration. If one splits ICE and Brown corpora data by genre, a better fit to the normal distribution emerges, if one splits by subgenres, typically a good fit arises, although the data are already rather sparse. The two subgenres *human science* and *natural science* of the scientific section are given below, in Figure 6 for Brown and in Figure 7 for ICE.
Figure 5. Passive percentages in the Brown family per file, when constraining the envelope of variation to alternatable variants

Figure 6. Two subgenres inside section J. vertical bars=HUM(ainties) and diagonal bars=NAT(ural science)
Regression models including subgenre and region report that region is significant, both in Brown and ICE, at least for the scientific section. As these data are normally distributed, one can now conclude that in the scientific section, there is significant regional variation between British and American English, as far as the LOB and ICE corpora are representative of the varieties in question. To summarize: In order to detect significant differences between varieties, it is important to include all relevant factors and model speaker choice as envelope of variation, i.e., a choice between two variants.

Passive versus active voice is a prime example of syntactic variation. In syntactic alternations in general, it is assumed that speakers and writers have a largely free choice between two alternatives, and hence they are good candidates for studies using an envelope of variation. Alternations are defined as syntactic variations which exploit the fact that “language can provide different syntactic means to express the same propositional content, to convey the same information” (Kreyer 2010: 169). Alternations have been extensively researched: Levin (1993) distinguishes 79 different alternations; some of the best investigated alternations include the passive voice alternation (e.g. Leech et al. 2009; Seoane 2009; Hundt et al. 2016); the dative shift (e.g. Mukherjee 2005; Bresnan & Nikitina 2009); and the Saxon Genitive alternation (e.g. Jucker 1993; Rosenbach 2002; Kreyer 2003; Röthlisberger & Schneider 2013).

For instance, Bresnan et al. (2007) use a logistic regression model to predict the dative alternation using a rich set of features. They predict up to 95% of the
binary decisions, their model only has a 5% residual. Thus, using an envelope of variation leads to more reliable results than using frequencies, and is often a reasonable operationalization.

2.1.3 Binary local decisions

Let us now address the question up to which point the choices which speakers make can be modeled by binary and relatively local decisions. We will suggest that global models which sum across all interrelated choices may be a useful addition, as we discuss in the following.

The fact that the decision for the alternation depends on very many factors, and that each of them is also a decision depending on other factors, indicates that there is no single place where choices are made, or decisions are taken; instead the decisions are distributed, they are non-local. Sinclair (2008), for instance, suggests looking at phraseology in terms of different levels of choice. Even if decisions are non-local, modeling them as local can of course still serve as a useful operationalization, but there may be other operationalizations which are more appropriate for some tasks. Additionally, selected contextual features can be added to regression models to reduce locality. On the one hand, the context in regression models needs to stay limited for reasons of sparse data, while parsers respect the context of the entire sentence. On the other hand, important selected features can also go beyond what today’s parsers typically respect. Bresnan et al. (2007), for example, use structural parallelism in dialogue as a feature, which often goes far beyond the current sentence. Even parsing itself has been approximated by using local features on finite-state cascaded chunkers (cf., e.g., Buchholz 2002).

Furthermore, the facts that the choice is quite limited, that the residual in Bresnan et al. (2007) is small, and that the trodden paths of using customary sequences are hardly ever left are in line with Sinclair’s idiom principle (Sinclair 1991), to be discussed in Section 2.2.1. A small residual does not entail that the choice is always deterministic, some variation is free and should not be predicted deterministically. The “noisy” properties are modeled explicitly in Bresnan and Nikitina (2009) as random effects.

Alternations are typically modeled as binary, two semantically similar constructions are in competition. Again, this is a useful operationalization for some tasks, but a clear simplification. Lehmann and Schneider (2012), for instance, note that the dative shift often has a third variant, the benefactive with the preposition for:

(1) Budget, […], has already offered 20% discounts for customers willing to book on line. (USAT644213)
(2) [...] many tour operators offer 3% discounts to travelers who don’t use credit cards. (LATM99072636)

Some verbs, for example provide, illustrate the various forms the alternation can take. The double-NP construction is not the alternative form to an NP + of-PP construction, but to an NP + with-PP construction, or an NP + to-PP construction. The double-NP construction (e.g. 3) is rare, the BNC only contains a few dozen of double-NP constructions, about 4,000 with-PP constructions (e.g. 4), and about 2,000 to-PPs (e.g. 5).

(3) You provided him his death, others have provided him a grave. (BNC-Wri K8S)

(4) The forwards played extremely well as a unit, driving in unison and providing their backs with good ball. (BNC-Wri K5A)

(5) Salespeople may also be called upon to provide after-sales service to customers. (BNC-Wri K94)

When including the benefactive construction, (e.g. bake Mary a cake vs. bake a cake for Mary) the alleged binariness generally collapses. In those cases where both for-PP and to-PP are available, the meaning is often completely different (do something for someone vs. do something to someone). Arppe et al. (2010), in a deliberately provocative statement, even dismiss alternations as a fundamentally flawed concept:

Our focus on alternations is the result of theoretical heritage from generative syntax and a matter of methodological convenience. Most linguistic decisions that speakers make are more complex than binary choices … alternations are as simplistic and reductionistic as the theories of language that originally studied them. (Arppe et al. 2010: 12)

If speaker choices are not local but spread across the entire utterance, and if they are not binary, then global statistical language models apart from simple operationalizations like alternations may be useful additions to logistic regression of alternation. The generation of each word is a speaker’s choice, a choice which is highly dependent on previous choices. Computational linguistics has used statistical models in which every word in its context is a choice point. Such statistical models are discussed in the following section, particularly with respect to the question if they could provide us with adequate models of language processing.
2.2 Models for natural language processing

Computational linguistics is practically inseparable from the use of statistical models. The application-driven need for the processing of natural language for tasks like speech recognition, part-of-speech tagging, syntactic parsing, or machine translation has led to the development of various application-oriented statistical language models. The first of these models, and still many successful ones, are based on surface sequences of words, so-called N-grams, where N is the length of the sequence.

2.2.1 N-gram models and the idiom principle

A very simplistic language generation model is the monogram model, in which the choice of a user is to draw a word token from an urn of word types, the lexicon, according to their observed frequency, \( p(\text{token}) \). If the previous word or words are also considered, we have N-gram models, in which the drawing of a word (token\(_0\)) is conditioned on the \( N - 1 \) previous word tokens, e.g. for \( N = 3 \):

\[
p(\text{token}_0 | \text{token} - 1, \text{token} - 2)
\]

Such models, which only take the sequence of words into account, seem extremely simplistic if not completely inappropriate. They only respect words at the surface level, they do not have any knowledge of hierarchical structure, in other words, of syntax. It rightly seems questionable from the viewpoint of the syntactician whether these models can solve any task; but in fact they are surprisingly successful, for example for speech recognition, part-of-speech-tagging, or machine translation (Mariño et al. 2006; Federico & Cettolo 2007). In fact, it is very difficult to improve a machine translation system by adding syntactic knowledge (cf., e.g., Koehn & Hoang 2007 and Sennrich 2013).

Sinclair (1991) pointed out that language production largely consists of prefabricated phrases (the idiom principle), and that linguistic creativity, building new utterances freely from syntactic rules (the open choice or syntax principle), is relatively rare. The idiom principle states that “A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (Sinclair 1991: 110). The syntax principle, by contrast,

is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word, phrase, or clause), a large range of choice opens up and the only restraint is grammaticality.

(Sinclair 1991: 109–110)
In rewrite-rule grammars, each rule is local, independent of nodes higher up or further down in the tree, a ’slot and filler’ model independent of the lexical material occurring in the terminal nodes. In Sinclair’s words, “Any tree structure shows it clearly: the nodes on the tree are choice points” (Sinclair 1991: 110).

In a radical version of the idiom principle, there are only surface sequences of words, and no syntactic abstractions: N-gram models are such a version. In a radical version of the syntax principle, all syntactic rules would be equally likely, irrespective of individual lexical items. There would either be no syntactic ambiguity, or no way to resolve ambiguity. The proportion of text consisting of phrases has been estimated in several studies. Pawley and Syder (1983) and Mel’čuk (1998) estimate the ratio of phraseological units to individual words at 10:1. Altenberg and Tapper (1998) estimate that up to 80% of the words in a corpus are part of a recurring sequence. All of these studies suggest that the idiom principle is the dominating force, and that an N-gram model is not a simplistic, but a fairly accurate model. Psycholinguistically, Tomasello (2000) shows that in language acquisition, formulaic language precedes the use of creative constructions, which gives primacy to the idiom principle. The chunking of sequential experiences is a process known in all cognitive systems (Newell 1990:7).

Wray (2002: 28) lists several studies which present surprisingly different counts for the ratio of phraseological units to individual words: between 4% to 80%, depending on what is included within the scope of phraseology. The largely differing counts indicate that recurrent phrases, idiomaticity, sequentiality and multi-wordedness are highly gradient concepts, and thus ask for a gradient measure.

Such a gradient measure of the influence of the idiom principle has been suggested by Levy and Jaeger (2007). The probability of the following word given the \( n \) previous words is called surprisal. It is an information-theoretic measure, and it assesses the entire gradience from idiom to rarity. Theoretically, surprisal can take the entire previous context of a discourse into consideration. For practical purposes however, going back more than a few words is hardly useful and, as corpora are limited, would inevitably lead to forbidding sparse data problems. In this study, we use bi-gram surprisal, which is defined as follows:

\[
(1) \quad \text{Bigram surprisal} = \log \frac{1}{p(w_1)} + \log \frac{1}{p(w_2 | w_1)}
\]

Surprisal is the logarithmic version of the probability of seeing word \( w_1 \) linearly combined with the probability of the transition to the next word, \( w_2 \). It is a N-gram language generation model, a model of predictability. Predictability is a key feature of multi-wordedness (e.g. Siyanova-Chanturia & Martinez 2014). This is an important point as numerous studies have shown that there is a strong correlation
between a word’s predictability in a given context and the ease with which it is comprehended (e.g. Seidenberg & MacDonald 1999).

The postulation that multi-wordedness is inherently gradient and measurable by frequencies and probability permeates usage-based linguistics.

As a particular string grows more frequent, it comes to be processed as a unit rather than through its individual parts. As it is accessed more and more as a unit, it grows autonomous from the construction that originally gave rise to it.

(Bybee 2006: 720)

Surprisal allows one to measure the competition between the idiom and syntax principles as follows. Language dominated by the idiom principle has low surprisal, is easy to process as its continuation is expected, but contains very little information. Language which makes maximal use of syntactic creativity can compress a lot of information into few words, but makes it very hard for readers or listeners to follow, and, as surprisal is very high, the continuation of the utterance is very hard to predict. According to Levy & Jaeger (2007), successful communication needs to strike a balance between the two: surprisal should stay constant across the entire text. This they call the principle of uniform information density (UID). “UID can be seen as minimizing comprehension difficulty” (Levy & Jaeger 2007: 850). So perhaps this is how we psycholinguistically form sentences: in a tug-of-war between the idiom and the syntax principle, between fixedness vs. expressiveness?

Further, there is a correlation between UID and the Grice’s cooperative principle. It can even be seen as an operationalization of the maxim of manner (Grice 1975):

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief.
4. Be orderly.

An abundance of areas of low surprisal flouts maxim 3, while areas of high surprisal flout maxim 1, and as we will argue in Section 5, maxim 2. Maxim 4 can either be seen as a consequence of the other maxims, or as a reminder that language offers word ordering options such as alternations.

Sperber and Wilson (1995, 2002) further elaborate on the Gricean cooperative principle in their approach called relevance theory. Relevance is seen as a function of processing effort and positive cognitive effect. Surprisal is a model of processing effort and positive cognitive effect at the same time.

Surprisal is a model of processing effort because idiomatic, expected continuations are easier to process: “Virtually every study, using a variety of research methodologies, shows that formulaic language holds a processing advantage over
nonformulaic language for native speakers” (Conklin & Schmitt 2012: 56). Low surprisal leads to low processing effort, and high surprisal to high processing effort. Positive cognitive effect can be equated to the amount of information that is transmitted. At least in the sense of information theory, where information is seen as the amount of unexpectedness, surprisal measures information. We thus need to expect that dense and scientific texts have consistently higher surprisal (as we will see in Section 3.1), as more information needs to be transmitted in relatively few words. However, for a given amount of information, relevance theory predicts that the least processing-intense ordering will be chosen by a native speaker:

In relevance-theoretic terms, other things being equal, the greater the PROCESSING EFFORT required, the less relevant the input will be. Thus, RELEVANCE may be assessed in terms of cognitive effects and processing effort:

Relevance of an input to an individual
a. Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.

b. Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time. (Sperber & Wilson 2002: 252–253)

In these terms, UID can be seen as an application of relevance theory: for a given idea to transmit, keep the processing effort, as measured by surprisal, as low as possible, by choosing an optimal sequence, which distributes the surprisal equally across the sentence, resulting in a balance between formulaicity and information load, avoiding both extremes of very high or very low surprisal, but forming a normal distribution of surprisal.

In order to illustrate this we measure surprisal at the level of word sequences, using bigrams, on selected BNC genres, namely spoken demographic and natural science, in Section 4.1. We also compare native language and learner language, on the learner corpora presented in Section 3.1. Learner language allows us to test Pawley and Syder (1983)’s hypothesis that native speakers know best how to balance fixedness vs. expressiveness, and adhere better to UID and relevance theory.

We are aware that, although N-grams are surprisingly realistic models of language processing, they are unable to model cognitive efforts such as inferencing in a given situation, or morphological or syntactic expectations. We will address at least the latter in the following subsection.

2.2.2 Syntactic models: Distributed interdependent decisions

Despite their success, N-gram and idiom principle models are of course insufficient, as they fail to take any hierarchical structural information (which one
usually calls syntax) into account. Syntactic parsers are language models which take syntactic structure into account. From a natural language processing (NLP) viewpoint, they allow us to detect the syntactic structure of a given sentence. From a language processing viewpoint, they also allow us to express how likely the mapping between a given sentence and its syntactic structure is. For almost every surface sentence many syntactic structures are possible and hence the problem of ambiguity arises in NLP, as we discuss first. Then we discuss a frequently used solution to the problem, which consists of a statistical model that takes lexical interactions and preferences into account. This statistical performance model turns out to be cognitively plausible, and thus offers a range of applications to psycholinguistic language processing research. In the last subsection, we argue that syntactic parsers are global processing models.

### 2.2.2.1 Ambiguity

In earliest approaches to automatic parsing, the amount of local and global ambiguity which syntax rules create was severely underestimated, often hundreds and thousands of analyses are possible for long sentences, which Wasow and Arnold (2003) summarize as follows:

Ambiguity is pervasive in natural languages. Indeed, early computational linguists were shocked to discover that their parsers typically found many more parses in any given string than native speakers realized were possible. Because ambiguity in one region of a sentence is typically independent of ambiguity in other regions, multiple local ambiguities create combinatorial explosions that can overwhelm parsers.

(Wasow & Arnold 2003: 146)

This is the other side of the coin of the recognition that the idiom principle is stronger than expected: while language generation is 1:few task (or at least fewer utterances than expected), syntactic analysis is a 1:many task: there are more possible syntactic analyses to a string than intuitively assumed. Syntactic ambiguity is the dark side of collocation.

### 2.2.2.2 The idiom and syntax principle in a tug-of-war

Sinclair (1991) observes that there are two opposing principles at work in language: the open-choice (or syntax) principle, in which syntactic rules dominate, and where lexical items can be placed freely in terminal node slots; and the idiom principle, in which frequently used word-sequences dominate, and syntactic rules (if they exist at all) are used only rarely and on larger units than local rewrite rules.

Sinclair’s syntax (open-choice) principle, in which competence grammar rules license possible analyses, and the idiom principle, in which performance probabilities rank them and constantly prune (filter) unlikely structures, is modeled as
a tug-of-war between these two principles by the parser. As the competence syntax delivers very many possible structures (it massively overgenerates), we need to use performance data for the disambiguation. For the performance disambiguation, we follow the approach of bi-lexical preferences (Collins 1999). The parser Pro3Gres (Schneider 2008) which we employ uses maximum-likelihood estimation (MLE) to estimate the probability of the dependency relation \( R \) at distance (in chunks) \( \text{dist} \), given the lexical head of the governor \( a \) and the lexical head of the dependent \( b \).

\[
(2) \quad p(R, \text{dist}|a, b) \propto p(R, a, b) \cdot p(\text{dist}|R) = \frac{f(R, a, b)}{f(\Sigma R, a, b)} \cdot \frac{f(R, \text{dist})}{f(R)}
\]

For example, the dependency rule for the object relation states that a noun preceded by a verb is licensed to be attached as object. The adjunct dependency rule is very similar, and often in conflict with the object rule. Say we have two adjacent words \( \text{dist} = 1 \) on the input stack at a certain point during the parsing process, a verb followed by a noun. If the verb is \( \text{eat} \) and the noun is \( \text{pizza} \), then the probability \( p(\text{obj}, 1 \mid \text{eat}, \text{pizza}) \) is high, while the adjunct probability is very low. If the verb is \( \text{eat} \) and the noun is \( \text{yesterday} \), the probability \( p(\text{adjunct}, 1 \mid \text{eat}, \text{yesterday}) \) is high, while the object probability is very low. The probability of each possible analysis for an entire sentence is the summation over the individual probabilities of each syntactic relation involved, according to formula (2). While these summed probabilities are intended to rank possible readings, they also constitute a score for each analysis.

For a syntax parsing model, the language analysis perspective is a more obvious choice than a language generation model (which we have used in Subsection 2.2.1): utterances that have already been generated are waiting to be analyzed. The language analysis probability \( p(\text{analysis} \mid \text{words}) \) and the language generation probability \( p(\text{words} \mid \text{analysis}) \) are related: if we use Maximum-Likelihood Estimation the relation is given by Bayes’ rule. This is in fact another difference between a parser model and a logistic regression model on an alternation (Bresnan et al. 2007): a parser aims to predict the best disambiguation from a listener’s perspective, an alternation regression aims to predict the variant from a language generation perspective.

Lexicalized statistical parsers do exactly what Sinclair (1991), Francis (1993), Hunston and Francis (2000) and Hoey (2005) predicted. They allow us to measure the strength of the idiom and syntax principles, patterns and rules, in interaction (or what psycholinguists refer to as routinization and chunking). Figure 8 gives an evaluation of the performance of the parser on the GREVAL corpus (Carroll et al. 2003). It compares a baseline model, which does not use statistical disambiguation (i.e. only the competence grammar is used, a pure open-choice principle model), a version which only uses the distance measure \( \text{dist} \) from equation (2),
a version which only uses lexical preferences \( p(R \mid a, b) \), full model without distance measure) and a version which uses both (the full model given in equation (2), open-choice principle ranked by idiom principle disambiguation). The evaluation is given using the standard metric F-score, which is the harmonic mean of precision and recall. We can see that lexical preferences and distance measure in combination reduce the error rate by about half, particularly the highly ambiguous prepositional phrase attachment relations (Noun-PP and Verb-PP) profit.

Other linguistically relevant factors can be added without limitations to this model. For example, in Schneider (2012) such added factors were semantic expectations. The underlying assumption for the parser to serve as psycholinguistic language model is proof by application: if the use of a feature improves parsing performance, then we can assume it to be (psycho-)linguistically influential.

The advantage of models is, in general terms, that they can take several factors into account and make predictions. The advantage of using a parser as a processing model is that it can supplement local models with a global model, and combine a multitude of local models in their interaction. A broad-coverage syntactic parser can serve as a language processing model for the following reasons: (1) It predicts attachment decisions based on grammar rules and lexical preferences. (2) Its statistical model can be extended by semantic and discourse-level factors. (3) It learns from real-word data, such as the syntactically annotated Penn Treebank (Marcus et al. 1993). (4) It assigns higher probabilities to entrenched structures, because they are expected.
2.2.2.3 Cognitive plausibility. While expected patterns obtain higher scores, unexpected structures obtain lower scores, due to the fact that the words as they are found in combination do not clearly point to a single analysis. This may be due to highly creative language use, due to unidiomatic combinations, or due to grammatical errors. Keller (2003) shows that there is a strong correlation between ungrammatical input and parser scores. Keller (2010) discusses the use of broad-coverage robust parsers as cognitively plausible models. However, approaches using global models on psycholinguistic tasks are still rare. Borensztajn, Zuidema and Bod (2009) have measured L1 language acquisition complexity using a Data-Oriented Parser (DOP) model (Bod et al. 2003), and Demberg, Keller and Koller (2013) use a similar Tree-Adjoining Grammar (TAG) approach, while Green (2014) uses a phrase structure and a dependency parser. They all use parser scores as well as lexical and structural surprisal to calculate global and local psycholinguistic processing difficulty. As language variation aims to explain reasons why sentences are rendered in specific ways, there is an enormous potential for using parsers as psycholinguistic language models in order to predict ambiguity and variation and to measure syntactic surprisal, and finally also to feed back the obtained knowledge into the parser. The parser score (formula (2)), which sums probabilities for a relation given the lexical participants over the whole sentence, is a measure of syntactic entrenchment, modeling syntactic expectations. More directly even, it is a measure of ambiguity: how clearly do the lexical participants point to a certain syntactic analysis, given past experience? This characteristic allows us to explore the role of ambiguity particularly well.

Green (2014) describes the correlation of parser entropy to text readability for human comprehenders concerning one specific type of ambiguity. He defines syntactic surprisal as the parser score at a given position divided by the parser score at an earlier position. The parser score of a sentence is thus a summation of the syntactic surprisal across the entire sentence.

2.2.2.4 Model parameters. Since the parser presents a language model to us, we can apply it in many different ways apart from measuring the signal (reporting significant differences in frequency), as one typically does in descriptive linguistics. We can modify model parameters, and use specific textual sources as input.

The model parameters indicated in Figure 9 can be changed minimally in order to describe the roles of ambiguity and disambiguation, formulaic language, language learner failures, readability, alternations, argument structure and lexical priming: on the parser’s input side, we can change the competence grammar, add more local models (e.g. Schneider 2012), vary the input by issuing L2 real world failure data or by forcing random alternations (voice, dative shift, synonyms, Saxon Genitive etc.), or vary the statistical model. Our research on issuing L2 real
Gerold Schneider and Gintarė Grigonytė

world failure data to the parser is described in Section 4, and forced input variation in Section 5.

2.2.2.5 Local and Global Models in Interaction. Since a parser is a global syntactic model, it gives us the right tool, for example, to punish a nested Saxon Genitive, to give high or low scores to analyses that are supported by, or unlikely, due to lexical preferences. Local models may miss the parsing context. We have discussed in Section 2.1.2 that Bresnan et al. (2007) use a logistic regression model to predict speaker choice in the dative alternation using a rich set of factors, and that the prediction accuracy of up to 95% of the binary speaker choices constitutes an excellent model. We have then pointed out that global models may be a useful addition due to the non-binariness of the choice, and due to the fact that the decisions may be interdependent and spread across the entire sentence. Examples of interdependence include persistence (Szmrecsányi 2006) or the Saxon Genitive, where nested’s constructions are extremely rare. Also, alternations are highly interdependent, as Arppe et al. (2010) point out. Syntactic relations are a prime example of highly interdependent factors: if a noun is subject, it cannot be an object, if a PP is attached to a verb it cannot be attached to a noun, and correct disambiguation depends on other disambiguations.

Speaker choices may be distributed across the selection of features and their interactions, and the interaction between the syntax and the idiom principle. In order to model psycholinguistic decisions, it can be helpful to use a global model which takes these interactions into account, combines and weighs local models, and uses the parsing context. This is precisely what a parser does. Parsers are not

**Figure 9.** Selected input (arrows to parser) and output (arrows from parser) parameters of the parser as a language processing model.
blackboxes which are difficult to interpret, but they have been carefully designed and evaluated to make maximally similar predictions to that of a native speaker in syntactic disambiguation tasks. The performance of a parser is a direct measure of how successfully all relevant factors are taken into consideration for this task. In other words: the better performing a syntactic parser is, the more adequate it is as a psycholinguistic processing model. One of the arenas in which this plays out is in the comprehensibility of learner language. We will apply a parser as processing model to learner language in Section 4. Learner language as a linguistic resource is introduced in the following section.

2.3 L1 and L2 data

According to Pawley and Syder (1983: 193), native speakers (L1) know best how to play the game of fixedness vs. expressiveness: “native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did do so they would not be accepted as exhibiting nativelike control of the language.” They hypothesized that one major reason for the reduced formulaicity of learner language (L2) is lack of input. Subsequent research showed that they were right:

Virtually every study, using a variety of research methodologies, shows that formulaic language holds a processing advantage over nonformulaic language for native speakers. However, for nonnatives, this is often not the case, although higher proficiency levels increase the chances of also enjoying this advantage. The crucial role of frequency in processing clearly applies not only to individual words but also to formulaic sequences. It appears that frequency of exposure is a key aspect of learning formulaic sequences. (Conklin & Schmitt 2012: 56)

Millar (2011) investigates native speakers’ processing of learner collocations which deviate from target-language norms. Results show that such deviations are indeed associated with an increased and sustained processing load, as hypothesized by Pawley and Syder (1983).

The close correspondence between frequency and routinization has contributed to the development of an approach to grammar in which frequency plays a central role and in which usage shapes grammar: so-called usage-based theories. Bybee (2007) summarizes as follows:

Frequency is not just a result of grammaticization, it is also a primary contributor to the process, an active force in instigating the changes that occur in grammaticization … I will argue for a new definition of grammaticization, one which recognizes the crucial role of repetition in grammaticization and characterizes it as the
process by which a frequently-used sequence of words or morphemes becomes automated as a single processing unit. (Bybee 2007: 337)

When native speakers construct sentences they employ information structure, word order, alternations (Levin 1993), choice of synonyms and register as subtle operations (Pawley & Syder 1983). Although grammatical variation seems abundant (cf., e.g., Rohdenburg & Mondorf 2003) it is severely restricted by complex and interacting factors. Sentences are rendered in the way they are due to many complex and interacting factors, and even subtle failures increase the processing load, both for the human parser and, as we will see, for an automatic parser. Language learners differ from native speakers since they often exhibit such failures. L2 corpora allow us to investigate real-world failures.

3. Data and methodology

In the following we will introduce the data and the methodology employed in our current study. The first subsection describes the learner corpora that we use, the second explains our method.

3.1 Data

We use the error-corrected Japanese Learner English (JLE) Corpus by the NICT for most of our investigations (Izumi et al. 2005). The corpus consists of spoken data from exam interviews and contains around 120,000 pairs of sentences, consisting of an original sentence produced by a language learner and a corrected sentence produced by a language teacher. The NICT-JLE corpus is thus a fully parallel corpus, controlled perfectly for content and topic, and only differing in original expression versus teacher-corrected expression, in those positions where a language teacher deemed a correction necessary.

In order to test proficiency levels and to investigate written data, we further use CEEAUS, the Corpus of English Essays Written by Asian University Students (Ishikawa 2009), 2012 Version. This corpus is not error-corrected, but the content is controlled by only using two essay topics, and the data are annotated for four proficiency levels of learners.

2. CEEAUS is now a part of the ICNALE project.
3.2 Surprisal and UID

To test how far UID (Section 2.2.1) holds, we measure surprisal at the level of word sequences, using bigrams, in the spoken demographic and natural science BNC genres (Section 4.1). We then compare native language and learner language in the learner corpora presented in Section 3.1. We thus test whether Pawley and Syder (1983)’s hypothesis holds, i.e. whether native speakers exhibit a more optimal balance between fixedness vs. expressiveness, and abide better to UID and relevance theory. Surprisal is a sequence-based language model, from which UID is derived as a consequence of the tug-of-war between keeping the processing burden low (low surprisal) while transmitting as much information as possible (high surprisal). UID is a principle rather than a model and we measure the degree to which language fits this principle.

3.3 High levels of residuals and low model fit of parsers as indicator

Syntactic parsers are language models designed to disambiguate real-world sentences. Models can be used in a variety of ways: they allow one to describe correlations, to predict outcomes, to separate significant differences (signal) from random fluctuations (noise), or to test how well a set of data fits a model. Model fit, on the one hand, refers to the accuracy of the model in predicting the data (for parsing the prediction is the syntactic analysis). On the other hand, model fit also refers to the confidence with which the model makes its predictions.

We evaluate the accuracy of our parser model in Sections 2.2.2 (Figure 8) and 4.2.1. If a model generally achieves reasonable accuracy, tokens or sets where the model has low confidence in its decision can be considered data which deviate from the standard the model has learnt its patterns from.

Our use of learner data (Section 4) and systematically forced input variation (Section 5.3) employs models as described immediately above: we expect learner data to fit the models trained on native speakers less well, to have lower model fit, higher residuals, lower parser scores, and we observe these differences. This is a method which is typically used for the detection of errors, anomaly, novelty and outliers in the data. Aggarval, for instance, states: “Virtually all outlier detection algorithms create a model of the normal patterns in the data, and then compute an outlier score of a given data point on the basis of the deviations from these patterns” (Aggarval 2013: 6). The model of the normal patterns in the data is the syntactic parser which has been trained on L1 language use. The outlier score is then the difference between the parser scores when applying the model to L2 data and to L1 data, or original L2 data and grammatically corrected L2 data. Probability-based scores, which are originally intended for disambiguation and ranking of
parsing candidates, can be used as a measure of the confidence with which a model takes its decisions. A high parser score indicates that the utterance matches the model’s expectation, that the found combination of lexical items strongly point to a certain analysis. A low parser score, by contrast, indicates that the utterance is unexpected by the model and that the parser cannot map it well to any known syntactic analysis.

Keller (2003) confirms these assumptions by showing that there can be a strong correlation between grammaticality and parser scores. In German, “SOV is generally regarded as the basic word order for subordinate clauses. Verb initial orders are regarded as ungrammatical.” Keller (2003: 646) compared human parser magnitude estimation scores against automatic parser probability, and reports a strong correlation.

4. Results: Two language processing models

In this section, we present and interpret the results obtained by applying the two language models introduced above to various English genres and to a comparison of native language and learner language, as well as learner language levels.

4.1 Surprisal at the level of word sequences

In accord with Pawley and Syder (1983) we expect Learner English to show violations of UID and higher surprisal, as discussed in Section 2.3. We have introduced surprisal in Section 2.1, and motivated our use of a learner corpus in Section 2.3, and introduced NICT-JLE in Section 3.1. In the following, we compare bigram surprisal between original (yellow bars) and corrected (green bars) learner English (Figure 10). The differences are small, but statistically highly significant (Welch Two Sample T-test, p < 2.2E-16). The mean bigram surprisal of the original is 11.73, compared to 11.55 for the corrected utterance. The uncorrected text has considerably more areas of high surprisal (16–19, to the right of the graph) and considerably fewer areas of low to medium surprisal (6–12), which also leads to a flatter normal distribution for the original learner language, showing that it adheres less to the principle of UID (Levy & Jaeger 2007). The bigram surprisal of the original utterances have a standard deviation of 3.48, compared to 3.36 for the corrected utterances.
If UID holds, we expect to see a balance between expressivity (high surprisal) and efficiency (low surprisal), leading speakers to choose, for whichever message they want to convey, sequences of words at an intermediate level of surprisal as often as possible, thus abiding by the Gricean Cooperative Principle (Grice 1975) and Relevance Theory (Sperber & Wilson 1995, 2002) as discussed in Section 2.2.1. Some bigrams with high surprisal correspond to typical learner errors, such as agreement errors or unusual word order. In fact, surprisal and related measures are also used as an important factor in the computational linguistic task of error detection, for example in the CoNLL 2014 Shared Task on Grammatical Error Detection In Learner Language (Ng et al. 2014).

Figure 10 also shows a considerable amount of fluctuations which may point to data sparseness, also there were unseen bigrams (bigrams which occur in JLE, but not in the BNC were given a surprisal bigram value of 19). In order to test UID on native speakers, we have also measured surprisal in the spontaneous speech section (spoken demographic) of the BNC, and indeed obtained the smoother distribution of the diagonal bars given in Figure 11, which better approximates the normal distribution postulated by UID and relevance theory.

We have mentioned that surprisal depends on the information to be conveyed, that an utterance is optimally relevant (in the sense of Relevance Theory) if its surprisal is lowest, all other factors being equal. We have also measured surprisal in the BNC natural science section. High surprisal is very frequent, as shown in Figure 11, which contrasts the pure science (yellow bars) and spoken demographic (green bars) sections.
We can glean several indications from these results. First, UID could possibly be a planning help for language production under time constraints instead of a mechanism to reduce ambiguity for the listener. Second, as surprisal depends on the topic and genre, we need to control for both as much as possible if we want to compare how much UID is observed in different texts. These different texts should be comparable, or ideally parallel corpora. The NICT-JLE corpus, which we introduced in Section 3.1, is such a parallel corpus. Except for correction of learner errors, the original and the corrected text are identical. Third, we need a model which respects Sinclair’s open-choice principle in addition to a word-level surprisal model. To address this, we describe the use of a syntactic parser as a model in the following section.

4.2 Syntactic parser as a processing model

In Section 2.2.2, we have described syntactic parsers as a model for the tug-of-war between the syntax principle, in which a *competence* grammar licenses possible analyses, and the idiom principle, in which *performance* preferences rank them and constantly prune (filter) unlikely structures. We now apply the parser to learner language.
4.2.1 Parser accuracy

For our investigation in this section, we have manually annotated just over 500 syntactic relations from random sentences of the NICT-JLE corpus, including both original sentences and their corresponding corrected sentences.

Figure 12. Parser Accuracy on >500 relations from original and corresponding corrected sentences in the NICT-JLE corpus

Figure 12 shows that the error rate has significantly decreased on corrected text, it has dropped almost by half. As expected, the parser performs less well on the original learner utterances, which contain many errors, but better on the same learner utterances after many of the mistakes have been corrected. Compared to the evaluation in Section 2.2.2.2 (Figure 8) the accuracy of the corrected learner data is considerably higher than on the general newspaper text of the GREVAL corpus, as the sentences are typically simpler.

An illustrative example is given by Figures 13 (original) and 14 (corrected). The mistagging of play as a noun in Figure 13 by the automatic parser does not mean that a human parser would also make such an error. But the fact that the tagger, which is trained on large amounts of real-world context data, suggests a noun here, indicates that a verb reading is surprising in this context based on previous experience, and human readers or listeners may show slightly increased processing load or minimal delays in comprehension.
Figure 13. Example of automatic parse of an original learner utterance from the NICT-JLE corpus

Figure 14. Example of automatic parse of the corresponding corrected learner utterance from the NICT-JLE corpus

4.2.2 Parser model fit
As explained in sections 2.2.2 and 2.3, our hypothesis is that L2 utterances, particularly those produced by low proficiency-level speakers, do not fit the model very well – neither the human listener nor the computational parser model (which we use as language model) – and thus lead to lower parser scores and to increased processing times for human listeners (cf. Millar 2011 for human parsers; Keller 2003 for automatic parsers). As explained in Section 3.3, probability-based scores reflect decision confidence and model fit. A high parser score indicates that the utterance matches the expectations of the model, for example if entrenched language and a particular combination of lexical items unambiguously point to one analysis.
A low parser score indicates unexpected input, the parser model cannot map it well to any known syntactic analysis.

The examples in Table 1 compare parser scores of original and corrected sentences from the NICT-JLE corpus. In the last sentence the corrected version obtains a lower score, but this is partly due to the fact that scores depend on sentence length.

Table 1. Parser scores obtained from original (ORIG) and corrected (CORR) NICT-JLE sentences

<table>
<thead>
<tr>
<th>Version</th>
<th>Sentence</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIG</td>
<td>Usually, I go to the library, and I rent these books.</td>
<td>5054.31</td>
</tr>
<tr>
<td>CORR</td>
<td>Usually, I go to the library, and I borrow these books.</td>
<td>8956.83</td>
</tr>
<tr>
<td>ORIG</td>
<td>For example, at summer, I can enjoy the sea and breeze.</td>
<td>7186.86</td>
</tr>
<tr>
<td>CORR</td>
<td>For example, in summer, I can enjoy the sea and breeze.</td>
<td>8965.99</td>
</tr>
<tr>
<td>ORIG</td>
<td>so I will go to the Shibuya three o’clock, nannda, before Hachikomae.</td>
<td>176.172</td>
</tr>
<tr>
<td>CORR</td>
<td>so I will go to Shibuya at three o’clock, nannda, in front of Hachikomae.</td>
<td>12787.4</td>
</tr>
<tr>
<td>ORIG</td>
<td>The computer game is very violence in today, but I do n’t like it.</td>
<td>6570.44</td>
</tr>
<tr>
<td>CORR</td>
<td>Computer games are very violent today, but I do n’t like them.</td>
<td>161.75</td>
</tr>
</tbody>
</table>

In order to test if the uncorrected sentences generally correspond to lower parser scores, we have analyzed and compared parser scores of original and corrected sentences, split by sentence length. Results are given in Figure 15. The parser score, which expresses model fit, is consistently higher for the corrected utterances. On average, the parser score of the corrected sentence is 1.37 times higher. The processing of non-native-like utterances leads to increased processing difficulties for the automatic parser, as well as the human parser.

Furthermore, to test if similar differences can be observed in written texts, and to see if parser scores increase in line with the level of proficiency of a given L2 speaker, we have measured parser scores according to proficiency level in the CEEAUS corpus. The parser scores are given in Figure 16, using a logarithmic scale, comparing low, mid, and semi-upper to upper level speakers. The line for lower proficiency level finishes at 25, because there are no sentences that are longer than 25 chunks in the data. Our expectations are largely confirmed: in most cases there is a correlation between speaker proficiency level and parser scores. The correlation may be less clear because the corpus is not as strictly parallel as NICT; nevertheless, the essay topics across speaker levels were identical.
Concerning second language acquisition (Granger 2009; Ellis 2012) our data support the view that learners are guided less by idiomatic utterances than by creative language use. They may be influenced by L1 transfer, lower L2 levels in particular, but more research will be needed on this subject.

Figure 15. Parser score by sentence length, measured in chunks, comparing original and corrected utterances

Figure 16. Parser scores by speaker level, split by sentence length
5. Ambiguity

While non-native language leads to lower model fit for the automatic parser and an increased processing burden for the human parser, it is contested if the increase in processing costs can be explained by increased ambiguity, and if native-like language is characterized by a maximal avoidance of ambiguity. We first look at garden-path sentences, which are the tip of the iceberg on the scale of ambiguity for the human reader, and which are generally rare.

5.1 Garden-path sentences

The hypothesis that we avoid ambiguity seems intuitively convincing. Compare the (fabricated) sentences (6) and (7)

(6) I saw the flower Ø I like
(7) I saw the flower Ø pots like

While the zero relative (Ø) in (6) is frequent, in (7) it is typically avoided, i.e. an overt relative pronoun (e.g. that) is used: first, the collocation between flower and pot triggers a garden-path noun phrase to be constructed in (7), second, pronouns cannot be pre-modified, which means that no such ambiguity can exist in (6) – indeed zero relatives in front of pronouns are particularly frequent – and third, inanimate nouns such as pots are rarely subjects of an active verb of perception.

Jaeger (2010) investigates the related zero-form of clause-initial complementizers. Based on UID, he shows that in areas of high information density (i.e. high surprisal for the intended reading) zero-forms are less likely to be used than in areas of low density. Schneider et al. (2005) showed that in garden-path situations created by sentences like (7) a locally most plausible interpretation needs to be revised based on subsequent text data if a globally possible interpretation is to be found. An automatic parser, much like a human parser, is easily caught in a local maximum which may prevent the global maximum from being found: in example (7), a NP-internal noun-modification relation (between flower as a dependent and pot as governor) is much more probable than a relative clause relation with a zero-form pronoun (with flower as governor and pot as dependent). If aggressive pruning is used, unlikely alternatives are cut and the globally correct reading cannot be found. The resulting necessary re-analysis then is similar to the human backtracking that is typically observed in eye-tracking experiments (e.g. Meseguer et al. 2002).
5.2 Avoidance of ambiguity

While there seems to be agreement that garden-path sentences are avoided by native speakers, the question whether we generally use information structure, word and constituent order etc. to avoid all types of ambiguity is contested. Wasow & Arnold (2003: 147) have investigated PP-attachment ambiguity and state that “there is little need for speakers to employ constituency ordering as a way of avoiding ambiguity”. They have investigated whether the dative shift alternation is more probable in ambiguous situations. In particular, whether a relatively heavy NP (which is typically avoided as the first constituent in the double-object construction variant) may be preferred as first constituent if it helps to avoid ambiguity. They used the to-PP sentence (8) and the double-object sentence (9).

(8) a. ambiguous: The foundation gave Grant’s letters to Lincoln to a museum in Philadelphia
   b. non-ambiguous: The foundation gave Grant’s letters about Lincoln to a museum in Philadelphia

(9) a. The foundation gave a museum in Philadelphia Grant’s letters to Lincoln
   b. The foundation gave a museum in Philadelphia Grant’s letters about Lincoln.

As expected, acceptability judgements for (9a), which avoids the ambiguity, were higher than for (8a). But their language production experiments pointed to the opposite trend: given the preposition about, (9b) was produced more often. On the one hand this may indicate that language production is less aware of potential ambiguity than language reception, and that avoidance of ambiguity only applies to some phenomena (e.g. zero-relatives and zero complementizers) but not to others (e.g. PP-attachment ambiguities). As PP-ambiguities are so pervasive, humans easily use other resources such as discourse and world knowledge to disambiguate.

On the other hand, a few additional remarks are in order. First, while Grant’s letters to Lincoln and Grant’s letters about Lincoln have the same weight in terms of words, their weight differs in terms of letters and in terms of surprisal: Bigram surprisal for letters to is 12.1, while bigram surprisal for letters about is 14.9 (see Figure 11 for a comparison). If Grant’s letters about Lincoln has higher weight, then (8b) is dispreferred, which leads to higher frequency of (9b).

Second, the principle of end-weight (Behaghel 1930; Hawkins 1994; Wasow 1997), which is an uncontested and very strong factor for constituent ordering, can itself be explained as a method to minimize ambiguity. Long, complex constituents open more potential attachment sites. If they appear at or close to the end of the clause or sentence, these attachment sites are not available in right-branching languages like English. In left-branching languages like Japanese or Korean,
a partial mirror image is reported by Hawkins (1994: 66–67): heavy constituents tend to appear at the beginning of sentences. This can be easily accounted for by avoidance of ambiguity, but not by temporal linearization such as given before new.

Third, Wasow and Arnold (2003) have only investigated isolated phenomena, and, as they point out themselves, it is unclear how far their experiment can be interpreted.

Fourth, their experiments were made in a controlled laboratory setting, not using real-world utterances. In the production experiment, participants read e.g. the sentence

\[(10) \quad \text{A museum in Philadelphia received Grant’s letters to Lincoln from the foundation.}\]

In a further experiment, the participants were asked to answer the question What did the foundation do? It is very well possible that, by then, the phrase letters to Lincoln was enough routinized and remembered as a single unit, with the result that the potential ambiguity was no longer perceived by the participants. As Jaeger (2010) stresses, it is important to use real-world utterances in addition to laboratory experiments, which always include some component of the observer’s paradox:

The final advantage of naturally distributed data I wish to discuss is arguably the most crucial one. As more and more research finds that speakers and listeners are sensitive to probability distributions […], it may be necessary to revisit the standard assumption that experiments with balanced designs are the best way to study language production and comprehension. (Jaeger 2010: 50–51)

Fifth, Wasow and Arnold (2003) write that sentences of the ambiguous type like (8a) are found extremely rarely: they only retrieved three in a large corpus consisting of articles from The New York Times. However, the fact that they are so rare is probably due to a reason. Our hypothesis is that they are rare because they are avoided, as is the case for garden-path sentences. In statistical systems, lack of data or sparse data is always a problem. In order to see how these unavailable examples would look like, and what their surprisal is, we construct them using syntactic alternations in the following section.

5.3 Forcing rare constituent order and alternative lexis

Garden-path sentences or the PP-ambiguity discussed in Section 5.2 are rare in real-world texts. Their rarity may be a meaningful sign of avoidance, or it may be a sparse data problem. In order to be able to observe them, we need to construct them in such a way that it does not alter the setting (i.e., introduces no observer’s paradox) but changes the data in syntactically permissible ways. As we treat
a parser as a syntactic model, we now apply an extension which forces random
alternations on the input text after parsing. We will refer to this extension as alter-
nator. In Sinclair’s terms, we are breaking the collaboration of the idiom and the
syntax principle and throw an exclusively syntax-driven spanner into the works.
The original (partial) sentence (11) after being subjected to a random Saxon geni-
tive alternation, which is permissible for benefactors, is turned into (12).

(11) …, Caldwell is expected to become a campaign coordinator for Byrd.
(12) …, Caldwell is expected to become a Byrd’s campaign coordinator.

Replacing a word with a synonym also constitutes a speaker choice that may be
called lexical alternation. We also apply a parser extension which performs ran-
dom synonym replacements, which we call synonymizer. After its application, the
original sentence (13) is turned into (14).

(13) Goodis voiced his objection before City Council’s Finance Committee.
(14) Goodis sounded his protest before City Council’s Finance Commission.

The alternator and the synonymizer can be thought of as artificial but systematic
L2 speakers, mechanical language learners who produce syntactically correct ut-
terances but consistently fail on lexical and collocation tasks, i.e. they have excel-
lent syntax principle knowledge, but reduced idiom principle competence.

In order to illustrate this, we applied the alternator and synonymizer to the
500 sentence GREVAL corpus (Carroll et al. 2003), which contains newspaper
language. After the application of the alternator, most sentences appear less well-
formed than in the original, and bigram surprisal increases from a mean of 13.5
for the original to 13.7 for the alternated text. After the application of the synony-
mizer, sentences almost always seem to be worse, and the median for the parser
scores on the texts with synonyms is four times lower. A plot of the score for the
synonymized sentence divided by the original sentence is given in Figure 17. A
value of 1 (which is found in 29 sentences) means that both versions received the
same score, which was typically the case for very short sentences. The scale con-
tinues much further to the right than shown here. There are sentences which have
a score which is several hundred times higher in the original (for example because
the new lexis could not be connected to a complete parse any more).

The fact that the application of syntactically allowed alternation and synonym
replacement operations leads to higher surprisal and lower parser scores is an in-
dication that they are less optimal in the sense of Relevance Theory, that Relevance
Theory is observed in most cases, and that typically the ordering which is most
compatible with Grice’s second maxim is generated.
6. Conclusions

In this study we have shown some advantages of using language-processing models typically used in computational linguistics for descriptive, quantitative and cognitive linguistics. The use of multifactorial models and an appropriate envelope of variation leads to statistically more reliable results than significance testing. For example, the application of these insights has been discussed with respect to passives, showing that there is regional variation. However, the envelope of variation is still a coarse approximation, as speaker choices are distributed across entire stretches of discourse. Accordingly, we have argued for global language processing models as psycholinguistic approximations to human language use. Two global models have been introduced: surprisal, which is a surface-based word-sequence model, and a syntactic parser, which is a hierarchical model.

Learner language has been tested on these models in Section 4, using the uniform information density principle (UID) on surprisal, and model fit on syntactic parser models. Our hypothesis that learner language fits the models less well, as predicted by Pawley and Syder (1983), is confirmed: UID is better observed in the corrected L2 utterances. When applying the Pro3Gres parser as language model, corrected L2 has higher parser accuracy and better model fit than the original utterances, and L2 proficiency also positively correlates with parser scores. UID also holds better for oral registers. Parsing can indeed be conceptualized in terms of the tug-of-war between Sinclair (1991)'s idiom and syntax principles, while surface
surprisal is a pure idiom principle model. As human parsing difficulties and automatic parsing difficulties correlate, a type of syntactic parser is indeed a good candidate for a psycholinguistic model. We intend to investigate these phenomena in a psycholinguistic experimental setting, and to apply parser models which include more factors. While lexical interaction and the idiom principle indeed seem crucial, other factors, such as persistence, are not yet included in today’s parser models, but technology is advancing fast.

Finally, in Section 5, we counteracted the lack of garden-path sentences, whose rareness in real-world utterances leads to a gap in the spectrum of observations, by systematically applying forced syntactic alternations, which typically affect parser scores and human comprehension alike, supporting the claim of Relevance Theory that orderings are typically optimal.

References


Name index

A
Ädel, Annelie 201
Aijmer, Karin 177, 215

B
Biber, Douglas 122, 124, 126, 127, 130, 143, 146, 202
Bod, Rens 299
Bresnan, Joan 289, 290, 297, 300
Bybee, Joan 2, 62, 146, 153, 294, 301, 302

C
Chomsky, Noam 4
Coates, Jennifer 30, 146, 148, 152, 153
Collins, Peter 16, 24, 128, 142, 144, 176, 180, 182, 183, 189–193
Crystal, David 3, 143

D
Dolphyne, Florence Abena 144, 145

F
Fetzer, Anita 91
Flowerdew, Lynne 201

G
Gass, Susan 252n3, 253, 254
Givón, Talmy 127, 130, 131, 270
Granger, Sylviane 176n2, 218, 219n2, 242, 282, 310
Gries, Stefan T. 2, 5, 51, 121, 122, 154, 281, 283, 286

H
Huber, Magnus 142, 144, 145, 169
Huddleston, Rodney 126, 146
Hudson, Richard 67–69, 72n4, 77, 77n8, 107
Hundt, Marianne 3, 5, 119, 127, 131, 176, 177, 179, 180, 247, 284, 287, 289
Hyland, Fiona 202, 203
Hyland, Ken 201, 202

J
Jaeger, Tim Florian 282, 293, 294, 304, 311, 313

K
Kachru, Braj B. 3, 5, 189
Keller, Frank 299, 304, 308
Klein, Wolfgang 254–257, 260, 273–275
Kortmann, Bernd 35, 36, 40, 41, 124
Kreiter, Rolf 289
Krug, Manfred G. 2, 4, 6, 24, 36, 43, 48n10, 62, 67–73, 75, 75n7, 80n10, 85n11, 89n12, 91, 92, 92n13, 94, 96, 102, 103, 176, 181, 188

L
Leow, Ronald 252–254, 272, 273
Lim, Lisa 118, 119, 124, 129, 130
Long, Michael H. 15, 252, 253, 270, 272

M
Mair, Christian 61, 118, 124, 142, 143, 167, 169, 177, 178, 180
Milroy, Lesley 15, 21, 21n8, 178
Morrow, Daniel 259, 264, 270, 271, 275
Mukherjee, Joybrato 5, 176, 215, 247, 289

N
Nassaji, Hossein 252, 253
Nikitina, Tatiana 289, 290

P
Palmer, Frank R. 143, 146, 151
Plato 4
Poplack, Shana 152, 167
Pullum, Geoffrey K. 77, 77n8, 126, 127, 146

Q
Quirk, Randolph 72, 104n18, 125–127, 130, 143, 146

R
Robinson, Peter 253, 254
Rosa, Elena 253, 272, 273

S
Schmidt, Richard W. 253
Schneider, Edgar W. 3, 14, 37, 47n8, 118, 141–143, 145, 168, 175
Schneider, Gerold 2–5, 10, 282, 284, 289, 290, 297–299, 311
Schramm, Andreas 4, 5, 9, 256, 258, 259, 270, 271
Sinclair, John McHardy  122, 282, 290, 292, 293, 296, 297, 306, 314, 315
Sperber, Dan  294, 295, 305
Strevens, Peter  3, 228
Swain, Merrill  253, 262
Szmrecsanyi, Benedikt  35, 38, 40, 40n5, 42, 142, 143, 152, 161, 167, 168, 300

T
Tagliamonte, Sali A.  16, 20, 230, 12, 24, 25, 144, 151, 152, 167, 176
Tognini-Bonelli, Elena  2
Torres Cacoullos, Rena  142–144, 147, 151–153, 165, 167
Trudgill, Peter  13, 14, 55, 61, 177

V
VanPatten, Bill  253, 270, 272, 273
Verplaetse, Heidi  103, 103n17, 104

W
Walchli, Bernhard  38, 40
Wasow, Thomas  296, 312, 313
Williams, Jessica  253, 273
Wilson, Deirdre  294, 295, 305
Wittgenstein, Ludwig  4
Z
Zwaan, Rolf  258, 259, 264, 270, 275
Subject index

<table>
<thead>
<tr>
<th>#</th>
<th>0-state situation 1-state situation, 2-state situation, see aspect, situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>AA approach see aggregative analysis</td>
</tr>
<tr>
<td></td>
<td>active voice see voice</td>
</tr>
<tr>
<td></td>
<td>agent 68, 125, 258 see also agentivity by-131, 126</td>
</tr>
<tr>
<td></td>
<td>agentivity 147n4, 153, 162–164 see also agent</td>
</tr>
<tr>
<td></td>
<td>aggregative analysis 6, 37, 39, 40, 45, 47, 48, 61 see also phenogram</td>
</tr>
<tr>
<td></td>
<td>alternation 124, 282, 287, 289, 290, 291, 294, 297, 299, 300, 302, 312–314, 316 see also alternator</td>
</tr>
<tr>
<td></td>
<td>alternator 314 see also alternation</td>
</tr>
<tr>
<td></td>
<td>ambiguity 282, 293, 294, 296, 299, 311–313 AmE see American English</td>
</tr>
<tr>
<td>B</td>
<td>BAWE see British Academic Written English Corpus</td>
</tr>
<tr>
<td></td>
<td>BE-passive 125, 129, 130, 135, 136 behavioral coordination 14 BNC see British National Corpus</td>
</tr>
<tr>
<td></td>
<td>BrE see British English</td>
</tr>
<tr>
<td>C</td>
<td>Caribbean English 124, 128, 144 causal inferencing 251, 256, 258, 265, 268, 270, 272, 274, 275 CEEAUS see Corpus of English Essays Written by Asian University Students CLARIN see Common Language Resource and Technology Infrastructure</td>
</tr>
<tr>
<td></td>
<td>classroom setting 8, 9, 169, 205, 215, 216, 235, 245, 253 see also naturalistic input</td>
</tr>
<tr>
<td></td>
<td>cluster analysis 40, 40n5, 41 co-text 91, 92, 94, 96, 99, 101 cognitive linguistics 1, 4, 315 CollateX 204, 208 colloquialization 128, 134, 177, 178, 181, 193 Common Language Resource and Technology Infrastructure 209 computational linguistics 2, 3, 281–283, 291, 292, 305, 315 concept-oriented approach see meaning-oriented approach</td>
</tr>
<tr>
<td></td>
<td>convergence 14–17, 21, 25, 29–31, 55, 167, 234 core modal see modal</td>
</tr>
<tr>
<td></td>
<td>corpus linguistics 1–3, 5, 201 Corpus of English Essays Written by Asian University Students 302, 302n2, 309 Corpus of English in Finland 8, 179 see also English in Finland</td>
</tr>
<tr>
<td></td>
<td>Corpus of English in Sweden 8, 179 see also English in Sweden</td>
</tr>
<tr>
<td></td>
<td>Corpus of Global Web-based English 71, 118n2</td>
</tr>
<tr>
<td></td>
<td>Corpus of Spoken Ghanaian English 145, 146 corpus-driven approach 2, 121, 122 CS-GH see Corpus of Spoken Ghanaian English</td>
</tr>
<tr>
<td>D</td>
<td>data-driven approach 2, 38, 45, 204, 205 dative shift 290, 312</td>
</tr>
</tbody>
</table>
degree of exposure to naturalistic language 213, 217, 230, 243–245
descriptive linguistics 283, 299
diachronic 145, 169, 175, 176, 178–180, 185
see also synchronic
dialect contact 13–15, 21, 30, 117–119, 125, 137, 138
difference matrix see distance matrix
diffusion 178, 194
disambiguation 297, 300
discourse marker 9, 213–227, 229–247
(dis)similarity matrix see distance matrix
distance matrix 36, 39–42
E
education see pedagogy
EFL see English as a foreign language
ELF see English as a lingua franca
emergent modal see modal
EngGram see English Constraint Grammar parser
English as a foreign language 61, 178, 197, 216, 217, 219, 221, 228–230, 234–247
English as a lingua franca 175–194
English as a second language 9, 120, 121, 137, 216, 217, 219, 228–230, 234–247
see also English as a foreign language
English Constraint Grammar parser 203, 208, 209
English in Finland 179n4 see also Corpus of English in Finland
English in Sweden 5, 215, 228, 230, 233–235, 235n8, 242, 245
see also Corpus of English in Sweden
English as a native language 239–241, 244
ENL see English as a native language
ESL see English as a second language
extraterritorial conservativism 177
F
feedback see peer and instructor feedback
FIN-CE see Corpus of English in Finland
first language 7, 58, 61, 175, 201, 282
FLOB see Freiburg-LOB Corpus
fluency 226, 242, 269, 270
Freiburg-Brown Corpus 182, 182n6, 187
Freiburg-LOB Corpus 182, 182n6, 187
frequency 2, 5, 226, 301
Frown Corpus see Freiburg-Brown Corpus
FTE see future time marking
future time marking 142–144, 147–154, 156–169
G
garden-path sentence 311, 312, 313, 316
genre 177, 181–183, 185, 200, 283, 285
GET-passive 119, 125–138
Ghanaian English 8, 141–146, 148–151, 153–170
GhE see Ghanaian English
globalization 6, 37, 45, 53, 55, 56, 58, 61, 62, 176, 177, 185, 193
GloWbE see Corpus of Global Web-Based English
grammatical aspect see aspect
grammatical variable see variable
grammaticalization 130, 131, 180
Gricean cooperative principle 294
I
ICLE see International Corpus of Learner English
idiom principle 282, 290, 292–296, 300, 306
see also syntax principle
imperfective 257–259, 261, 262, 264, 268, 270, 272, 274, 275
see also perfective
institutionalized second-language 9, 216, 235, 245
intention 68, 92, 94, 147, 147n4, 149, 258
interlanguage 176n2, 215, 216, 252, 253, 260, 262
see also Louvain
International Database of Spoken English
International Corpus of English 3, 117n1, 133, 146, 170, 287, 289
International Corpus of Learner English 203, 248
J
Jamaican Creole 118, 124, 128, 129
Japanese Learner English Corpus 302
JLE Corpus see Japanese Learner English Corpus
L
L1 see first language
L2 see second language
language acquisition 4, 14, 168, 214, 217, 218, 244, 252n1, 254, 260, 262, 272, 273, 282, 293
see also second dialect acquisition, second language acquisition
Language Archive 208, 209
modal 8, 75, 75n7, 80, 92, 102, 142, 151, 169, 175–177, 180–194
auxiliary 80n10, 176
core 8, 180–191
descriptive 8, 180–194
semi- 137, 144, 150
moment-to-moment processing 258, 265–267, 269
mood 77, 85
morphosyntactic aspect see grammatical aspect
MUCH see Malmö University-Chalmers Corpus of Academic Writing as a Process
multilingual 177, 178, 179n4, 189
multiple regression see regression model
N
N-gram model 292–295
see also non-native speaker
nativization 37, 142, 145, 168
naturalistic input 9, 214, 217–219, 221, 227, 244, 245
see also classroom setting
NeighborNet see non-hierarchical phenogram
New English 141–144, 168, 170, 176
New Rhetoric approach 200
New Zealand English 15, 20, 24, 27
NICT-JLE see JLE corpus
NNS see non-native speaker
non-hierarchical phenogram see phenogram
non-native speaker 8, 9, 28, 29, 175, 177, 178, 183, 184, 186, 187, 194, 215, 216, 235, 237, 238, 241, 244, 246, 251, 258, 259, 261, 262, 267–274, 311
see also native speaker
NS see native speaker
O
obligation 16, 20, 23–26, 30, 31, 96, 97, 98, 107
verbs of 16, 20, 23–26, 31, 286
off-line processing 268, 271
ongoing grammatical change 176, 177, 193
open-choice principle see syntax principle
outlier 43, 45, 48, 303
overuse 119, 136, 180, 215
P
parser 3, 5, 10, 203, 208, 282, 290, 296–311, 314–316
automatic 5, 10, 203, 208, 296, 302, 307–309, 311, 316
human 10, 302, 304, 307–309, 311, 316
model 297, 303, 308, 309, 314, 316
see also regression model
syntactic 10, 282, 292, 296–298, 301, 303, 306, 314
passive voice see voice
see also BE-passive, GET-passive
pedagogy 185, 197–199, 202, 205, 209, 236
peer and instructor feedback 199, 200, 202–204
perfective 257–259, 261, 262, 265, 265n7, 267, 268, 272, 274, 275
see also imperfective
persistency 167, 300, 301, 316
phenogram 40, 41, 43, 47, 48
non-hierarchical 40, 41
phylogenetic network see non-hierarchical phenogram
positive cognitive effect 294, 295
post-colonial language 3, 14, 168
pragmatics 213, 215, 216, 254, 257, 273, 274
PRE see Puerto Rican English
processing see language processing
propagation 108
<table>
<thead>
<tr>
<th>Subject</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>psycholinguistics</td>
<td>10, 281, 282, 293, 294, 296, 298–301, 315, 316</td>
</tr>
<tr>
<td>Puerto Rican English</td>
<td>7, 36, 37, 37n2, 43, 47, 50–61</td>
</tr>
<tr>
<td>questionnaire data</td>
<td>36, 37, 39n4, 47, 48, 61</td>
</tr>
<tr>
<td>random effects model</td>
<td>see mixed effects regression model</td>
</tr>
<tr>
<td>regression model</td>
<td>2, 10, 20, 26, 49, 51, 52, 153n7, 154, 154n9, 283, 290, 297</td>
</tr>
<tr>
<td>see also parser model</td>
<td></td>
</tr>
<tr>
<td>linear</td>
<td>49, 51, 52</td>
</tr>
<tr>
<td>mixed effects</td>
<td>2, 49, 51, 154</td>
</tr>
<tr>
<td>multiple</td>
<td>20, 26</td>
</tr>
<tr>
<td>relevance theory</td>
<td>281, 294, 295, 314, 316</td>
</tr>
<tr>
<td>repetition</td>
<td>see persistence</td>
</tr>
<tr>
<td>second dialect acquisition</td>
<td>14</td>
</tr>
<tr>
<td>second language</td>
<td>4, 37n2, 168</td>
</tr>
<tr>
<td>see also second dialect acquisition</td>
<td></td>
</tr>
<tr>
<td>acquisition</td>
<td>4, 8, 119, 138, 168, 247, 251–254, 272, 272, 310</td>
</tr>
<tr>
<td>semi-modal</td>
<td>see modal sequence</td>
</tr>
<tr>
<td>sequence</td>
<td>94, 257, 292, 293, 295, 301, 304</td>
</tr>
<tr>
<td>short-term accommodation</td>
<td>18</td>
</tr>
<tr>
<td>short-term memory</td>
<td>see working memory</td>
</tr>
<tr>
<td>similarity matrix</td>
<td>see distance matrix</td>
</tr>
<tr>
<td>Singaporean English</td>
<td>7, 118–125, 128–138, 144, 216, 245, 246</td>
</tr>
<tr>
<td>colloquial</td>
<td>118, 124, 128–132, 136, 137</td>
</tr>
<tr>
<td>situation aspect</td>
<td>see aspect</td>
</tr>
<tr>
<td>SLA</td>
<td>see second language acquisition</td>
</tr>
<tr>
<td>social network</td>
<td>15, 17, 20–22, 26–28, 30, 178</td>
</tr>
<tr>
<td>approach</td>
<td>6, 13</td>
</tr>
<tr>
<td>sociolinguistics</td>
<td>154, 170, 176, 177, 253, 284, 285</td>
</tr>
<tr>
<td>spoken language</td>
<td>142, 150, 177, 180, 181, 184, 189, 215, 216, 228</td>
</tr>
<tr>
<td>see also written language</td>
<td></td>
</tr>
<tr>
<td>standard language</td>
<td>70, 71, 117, 129</td>
</tr>
<tr>
<td>English</td>
<td>118, 130, 142, 193</td>
</tr>
<tr>
<td>statistical model</td>
<td>2, 281–283, 291, 292, 296</td>
</tr>
<tr>
<td>substrate influence</td>
<td>117, 128–130, 138, 168</td>
</tr>
<tr>
<td>SWE-CE</td>
<td>see Corpus of English in Sweden</td>
</tr>
<tr>
<td>Sweden</td>
<td>see English in Sweden</td>
</tr>
<tr>
<td>synchronic</td>
<td>178, 185, 186</td>
</tr>
<tr>
<td>see also diachronic synonymizer</td>
<td>314</td>
</tr>
<tr>
<td>syntax principle</td>
<td>292–294, 296, 297, 306</td>
</tr>
<tr>
<td>see also idiom principle</td>
<td></td>
</tr>
<tr>
<td>tagging</td>
<td>203, 283, 292</td>
</tr>
<tr>
<td>temporal structure</td>
<td>251, 252</td>
</tr>
<tr>
<td>textual stratification</td>
<td>176</td>
</tr>
<tr>
<td>UID</td>
<td>see uniform information density principle</td>
</tr>
<tr>
<td>uniform information density principle</td>
<td>294, 295, 303–306, 315</td>
</tr>
<tr>
<td>variability</td>
<td>58, 178, 180, 189, 193, 194</td>
</tr>
<tr>
<td>variable</td>
<td>10, 26, 27, 28, 30, 31, 39, 49, 52, 141, 153n7, 154, 168, 170, 178, 214, 245, 246, 254, 264, 266–268, 281, 282</td>
</tr>
<tr>
<td>grammatical</td>
<td>16, 30, 31</td>
</tr>
<tr>
<td>linguistic</td>
<td>16, 45</td>
</tr>
<tr>
<td>phonological</td>
<td>15</td>
</tr>
<tr>
<td>predictor</td>
<td>27, 28, 49, 52, 282</td>
</tr>
<tr>
<td>sociolinguistic</td>
<td>169, 170</td>
</tr>
<tr>
<td>variationist linguistics</td>
<td>3, 154, 285</td>
</tr>
<tr>
<td>verbs of obligation</td>
<td>see obligation</td>
</tr>
<tr>
<td>vernacular</td>
<td>168, 180</td>
</tr>
<tr>
<td>Vienna-Oxford International Corpus of English</td>
<td>180, 183–185, 190</td>
</tr>
<tr>
<td>viewpoint aspect</td>
<td>see grammatical aspect</td>
</tr>
<tr>
<td>visualization</td>
<td>6, 36, 40–43, 48, 198</td>
</tr>
<tr>
<td>VOICE</td>
<td>see Vienna-Oxford International Corpus of English</td>
</tr>
<tr>
<td>voice</td>
<td>125, 287, 289</td>
</tr>
<tr>
<td>active</td>
<td>129</td>
</tr>
<tr>
<td>passive</td>
<td>125, 126, 128, 129, 286–289</td>
</tr>
<tr>
<td>volition</td>
<td>92, 94, 100–104</td>
</tr>
<tr>
<td>WM</td>
<td>see working memory</td>
</tr>
<tr>
<td>word completion</td>
<td>260, 265</td>
</tr>
<tr>
<td>see also long-term memory</td>
<td></td>
</tr>
<tr>
<td>World Englishes</td>
<td>3, 4, 7, 35, 69, 121, 168</td>
</tr>
<tr>
<td>writing process</td>
<td>8, 197–200, 202–205, 209</td>
</tr>
<tr>
<td>writing research</td>
<td>9, 197–205, 209</td>
</tr>
<tr>
<td>written language</td>
<td>8, 119, 127, 128, 134, 169, 180, 181, 184, 197, 199, 228</td>
</tr>
<tr>
<td>see also spoken language</td>
<td></td>
</tr>
</tbody>
</table>
This book aims at providing a cross-section of current developments in English linguistics, by tracing recent approaches to corpus linguistics and statistical methodology, by introducing new inter- and multidisciplinary refinements to empirical methodology, and by documenting the on-going emphasis shift within the discipline of English linguistics from the study of dominant language varieties to that of post-colonial, minority, non-standardised, learner and L2 varieties. Among the key focus areas that define research in the field of English linguistics today, this selection concentrates on four: corpus linguistics, English as a global language, cognitive linguistics, and second language acquisition. Most of the articles in this volume concentrate on at least two of these areas and at the same time bring in their own suggestions towards building bridges within and across sub-disciples of linguistics and beyond.