

Conversion in Asian Englishes

A usage-based account of the emergence
of new local norms

Stephanie Horch



Conversion in Asian Englishes

A usage-based account
of the emergence of new local norms

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
der Philologischen Fakultät
der Albert-Ludwigs-Universität
Freiburg i. Br.

vorgelegt von

Stephanie Horch
aus München

Wintersemester 2015/16

Erstgutachter: Prof. Dr. Dr. h.c. Christian Mair

Zweitgutachter: Prof. Dr. Martin Hilpert

Vorsitzender des Promotionsausschusses
der Gemeinsamen Kommission der
Philologischen, Philosophischen und Wirtschafts-
und Verhaltenswissenschaftlichen Fakultät: Prof. Dr. Hans-Helmuth Gander

Datum der Disputation: 06.04.2016

Contents

1	Introduction	1
1.1	Aims and scope of the study	3
1.1.1	Why investigate conversion?	3
1.1.2	A usage-based account of the deverbal converted noun construction	4
1.1.3	Varieties of Englishes	6
1.2	Research questions	27
2	Previous research	33
2.1	Defining conversion	33
2.1.1	Terminology	34
2.1.2	Productivity and constraints on conversion	37
2.1.3	Light-verb constructions	42
2.2	Conversion in the substrates	43
2.2.1	Conversion in Chinese dialects	44
2.2.2	Conversion in Malay	46
2.2.3	Conversion in Hindi	47
2.3	Conversion in World Englishes	47
2.4	Modeling frequency effects in language contact	50
2.5	Summary	54
3	Theoretical framework	55
3.1	Implementing usage-based modeling in word formation	55
3.1.1	Conversion versus derivation	55
3.1.2	Chunks and lexical diffusion	56
3.1.3	Processing conversion	57
3.2	A usage-based account of variety genesis	58
3.3	Summary	62

Contents

4	Data and methods	65
4.1	Corpora	65
4.1.1	<i>International Corpus of English</i>	66
4.1.2	<i>Corpus of Contemporary American English</i>	67
4.1.3	<i>Corpus of Global Web-based English</i>	69
4.1.4	Potential and limitations of corpora	72
4.2	Quantitative methods	80
4.2.1	Collocation analysis	81
4.2.2	Linear regression	82
4.2.3	Logistic regression	85
4.2.4	Linear and logistic regression with random effects	86
4.2.5	Frequency as a continuous predictor	87
4.2.6	Model criticism	89
4.3	Experimental methods	92
4.3.1	Acceptability judgment	92
4.3.2	Reaction time	94
4.3.3	Web-based experimentation	95
4.4	Summary: Integrating quantitative and qualitative analyses	101
5	Conversion as a productive process in US English	103
5.1	DISCONNECT VS. CONNECT	103
5.1.1	The 'rise' of DISCONNECT (N) and the blocking of CONNECT (N)	104
5.1.2	Semantic shift	110
5.1.3	Functional shift and syntactic variability	113
5.1.4	Emergence of the plural form	120
5.1.5	Register analysis	122
5.1.6	Interim summary	125
5.2	Further examples	126
5.2.1	DIVIDE	126
5.2.2	INVITE	129
5.2.3	PAY	131
5.3	Summary: Conversion in USE	132
6	A quantitative approach to conversion in World Englishes	135
6.1	Hypotheses	135
6.2	Corpus samples	137

6.3	Results and discussion	139
6.3.1	A ‘colonial’ model of conversion in Englishes	139
6.3.2	The globalized picture	147
6.3.3	Excursus: Refining the dataset	152
6.4	Summary	156
7	A qualitative approach to conversion in Asian Englishes	159
7.1	Transfer from the substratum in Hong Kong English conversion	159
7.1.1	Registers and formality	159
7.1.2	Conversion in ICE-HK	162
7.1.3	Syntactic contexts	166
7.2	Constraining transfer in nativization: Examples from Singapore English	169
7.2.1	Registers and formality	169
7.2.2	Conversion in ICE-SIN	171
7.2.3	Syntactic contexts	172
7.3	Liberal use of conversion in Indian English	174
7.3.1	Registers and formality	175
7.3.2	Conversion in ICE-IND	176
7.3.3	Syntactic contexts	177
7.4	Further observations	179
7.4.1	Lexicalized formations	179
7.4.2	Analogical formations	183
7.4.3	Light-verb frames	184
7.5	Locating conversion on the lexis-syntax continuum	187
7.6	Cross-varietal differences in register	190
7.7	Summary	191
8	Experimental validation of corpus results	195
8.1	Task 1: Rating task	196
8.2	Task 2: Maze task	197
8.3	Task 3: Background questionnaire	200
8.4	Hypotheses	200
8.5	Materials and design	201
8.5.1	Rating task	201
8.5.2	Maze task	205
8.5.3	Background questionnaire	207

Contents

8.6	Procedure	207
8.7	Pre-test	208
8.8	Participants	209
8.9	Results	212
8.9.1	Rating task	214
8.9.2	Maze task	226
8.10	Discussion and summary	242
9	Discussion and conclusion	245
9.1	V>N conversion in USE	245
9.2	V>N conversion in Asian varieties	248
9.3	The interplay of substrate influence and indigenization	250
9.4	Asian ESL varieties	257
9.5	Processing V>N conversion	259
9.6	A note on methodology	260
9.7	Conclusion	262
	Appendices	291
A	Transcription conventions	291
B	Development of DISCONNECT and CONNECT	292
C	Further candidates for conversion in USE	296
D	Logistic regression models for conversion in World Englishes	299
D.1	Additional coefficients for the ‘colonial’ model	299
D.2	Additional coefficients for the ‘global’ model	299
D.3	An alternative logistic regression model	300
D.4	Additional coefficients for the trimmed ‘global’ model	301
E	Experiment on conversion in World Englishes	302
E.1	List of stimuli for the rating task	302
E.2	List of stimuli for the maze task	304
E.3	Background questionnaire	308
E.4	The <i>QualityCrowd2</i> tool	309
E.5	Completion times	313
E.6	Metadata of participants	313
E.7	Additional coefficients for the rating model	315
E.8	Further analysis of the rating task	316
E.9	Residual diagnostics for the maze task data	317

E.10	First model fitted to the maze task data	318
E.11	Histogram of trimmed data set	319
E.12	Additional coefficients for the model fitted to the trimmed data set .	320
E.13	Further analyses of the maze task	320

Acknowledgments

This book is a revised version of my PhD thesis written between the years 2013 and 2016. Over the course of the last few years, many people have contributed to making this project a success. First and foremost, I thank my supervisors, Christian Mair and Martin Hilpert, for all their help, support, constructive criticism and encouragement. I could not have wished for better supervisors!

I further thank Hans-Jörg Schmid for encouraging me to venture into academia and to pursue a PhD.

For their valuable comments on my work I am particularly indebted to (in alphabetical order) Bao Zhiming, Thomas Hoffmann, Claudia Lange, Ute Römer, and the participants of the *Changing English* conference in June 2015 in Helsinki.

The quantitative part of the present study would not have seen the light of day without Christoph Wolk's and Göran Köber's invaluable help in coming to grips with the intricacies of programming and statistical analysis. Their patience and speed in answering my many emails is only paralleled by Clemens Horch, who I owe thanks for adapting the *QualityCrowd* software for the experiment as well as for his advice and help with layout.

Writing a PhD thesis was a challenging experience and part of my success in completing it is due to the wonderful company at the office. In particular I would like to thank all those who have become friends over the past years, namely (again in alphabetical order) Udo Baumann, Annette Fahrner, Katja Roller, Laura Terassa, Vanessa Tölke, and Martina Zier.

For the stimulating environment and the generous funding for workshops and conference trips I am greatly indebted to the Research Training Group 1624 "Frequency effects in language", funded by the *Deutsche Forschungsgemeinschaft* (DFG), and everyone who forms a part of it, especially Michael Schäfer, who time and again took organizational duties off my shoulders.

In addition, I am grateful to the many unnamed people who I have had the pleasure of working with, who have supported me, and who have assisted me with the completion of this PhD thesis.

I also wish to thank the very diligent proof-reading task force, consisting of Udo Baumann, Suzette Golden-Greenwood, Katja Roller, Mairi Sinclair, and Laura Terassa. Any remaining mistakes are, of course, my own.

Finally, last but certainly not least, I thank Clemens for his boundless optimism and never-failing support from the very first to the very last day of this endeavor.

1 Introduction

The motivation behind the present study on conversion in Asian varieties of English is twofold. The first aim is to envisage conversion from the World Englishes perspective, the second is to focus on conversion from a usage-based perspective. Conversion is the change of word class without overt morphological marking, exemplified in 1.1.

(1.1) There's a **disconnect** between reality and perception. (COCA-NEWS)

The study in hand is located within the framework of World Englishes. Previous studies on varieties of English have mostly focused on aspects that are easily accessible and immediately catch the linguist's attention (cf. Davies and Fuchs 2015b: 2), such as phonetic or grammatical variation. A further comparatively well-covered linguistic domain is the area of lexis, focusing mostly on borrowings from the substrate languages or hybrid formations (cf. e.g. Hashim and Leitner 2011; Siew Imm 2009 both for Malaysian English; Dako 2001 for Ghanaian English; Tent 2001 for Fiji English). Lexical variation, particularly between native varieties of English, has even found its way into school curricula and, subsequently, into school books for English as a foreign language,¹ while less obvious features pertaining to the lexical domain, such as minor word-formation processes, have not been as extensively explored. With the availability of large corpora, however, the analysis of minor, comparatively infrequent phenomena by means of collocations and n-gram analyses is now within reach, as demonstrated in e.g. Schilk (2011) or Gries and Mukherjee (2010). Nonetheless, the process of conversion is still rather uncharted territory. Baumgardner (1998: 229), in his study on word formation in Pakistani English, covers the topic of conversion but like many other studies on word formation he does not provide a detailed explanation as to why the formations observed in contact varieties "can also be found in native varieties of English, but not to such a degree". One reason why conversion is as of yet comparatively under-researched is its location at the notoriously difficult to access lexis-grammar interface. As Gries and Mukherjee (2010: 525–526) point out

¹Examples are the book series *Green Line New* and *English G21* for secondary education in Germany, which raise awareness for the topic of varieties of English as early as in grade 8 (Ashford et al. 2000 and Abbey et al. 2011, respectively).

1 Introduction

[i]n spite of a growing interest amongst a number of linguists, the lexis-grammar interface is still largely a blind spot in research into many postcolonial varieties of English. This has to do with the fact that at the lexico-grammatical level, e.g. with regard to collocations and verb-complementational patterns, differences between varieties of English are usually not categorical but quantitative in nature, so that large and representative corpora are needed to identify different trends and preferences across varieties of English.

The currently most extensively used corpus for research into World Englishes is the *International Corpus of English* (ICE). For every variety, ICE contains approximately one million words. While this is still a reasonable size when it comes to investigating grammatical patterns, the ICE sub-corpora fail to provide sufficient evidence for most word-formation processes (cf. e.g. Biermeier 2008: 198). Nelson (2004: 226) admits that

indeed lexical study has never been our [i.e. the ICE compilers'] primary objective. Since the project was first mooted by Greenbaum, our long-term aim has been to tag the corpora for parts of speech, and to parse each corpus syntactically, so that researchers can compare varieties of English at the level of syntax.

The issue of conversion in new varieties of English can thus be said to find itself in the proverbial ivory tower. This is due to the fact that, firstly, large corpora documenting subtle trends in usage have not been available and that, secondly, word formation, and particularly conversion, is caught in a limbo between grammar and lexis. None of the traditional approaches to either grammar or lexis cover conversion because it is not a prototypical part of grammar nor is it a prototypical word-formation process (cf. Bauer 2003: 124). Hence, investigating conversion in New Englishes is a much needed undertaking, particularly since new, larger corpora such as the *Corpus of Global Web-based English* (GloWbE, Davies 2013) facilitate investigations into infrequent phenomena by providing the required amounts of data.

The availability of large corpora provides the opportunity for quantitative analyses. Whereas previous studies in word formation in World Englishes (cf. e.g. Biermeier 2008) have oftentimes refrained from any statistical analysis beyond mere counting, larger amounts of data open up new possibilities in this respect. This complements the corpus-linguistic paradigm in that in-depth qualitative case studies can be combined with statistical profiling (a combination that Mair calls for already in 2007, cf. Mair 2007). This study consequently seeks to combine traditional case studies, also drawing on smaller, traditional corpora such as ICE, with statistical modeling on the basis of larger, less meticulously compiled corpora such as GloWbE.

A quantitative approach to language variation falls within a usage-based approach to language, which is the second framework of this study. The usage-based approach strives to understand how the usage frequency of language phenomena structurally influences e.g. language change, language processing, and language acquisition. Common methods often employed in this line of research include statistical modeling based on previous corpus analyses as well as the analysis of experimental data. This project aims to contribute to this line of research and extend it to the fields of word formation and of World Englishes. It aims to explore the role of frequency in shaping varieties of English, more precisely, to ascertain in how far frequencies of occurrence of diverse constructions (e.g. particular verbs, near synonyms) can influence the usage patterns of conversion in different varieties of English. A further goal is to explore whether frequency-related constraints that operate on native varieties of English also apply to the same degree to Asian varieties of English.

1.1 Aims and scope of the study

1.1.1 Why investigate conversion?

The phenomenon of conversion, i.e. the change of word class without morphological marking, is extremely frequent, not only in English but also in other languages around the world (cf. Štekauer et al. 2012: 309). For English, there are countless examples of conversion between nouns, verbs, adjectives, adverbs, and prepositions. The most productive direction of conversion is the derivation of a verb from a noun (e.g. *to google*, *to access*). Other languages also show conversion, especially those with little or no inflection such as Chinese (cf. chapter 2.2). Understanding how two different usage patterns of conversion interact in the formation of contact languages seems a rewarding undertaking in the field of World Englishes.

As has been mentioned already, conversion is also of interest because of its ambiguous status between grammar and lexis. This study adopts a Construction Grammar approach which overcomes this traditional opposition and in doing so helps to depict the phenomenon of conversion more precisely.

In this project, the direction of conversion from verb to noun is investigated. Since noun-to-verb conversion is the most frequent, i.e. most productive, type of conversion in English, it seems that it is also the most unconstrained direction of conversion. Verb-to-noun conversion is less frequent in English and—as it seems—more constrained. A moderate level of productivity in native varieties of English is ideal for a comparison with New Englishes. If a substrate language were to reduce or foster the productivity of verb-to-noun conversion,

1 Introduction

this would be easily visible, whereas in the case of noun-to-verb conversion productivity might always be high and, thus, could not unequivocally be traced back to the influence of a substratum or other factors inherent in the contact situation.

Within the area of verb-to-noun conversion, this study deals with those instances that challenge the idea of a blocking constraint. The blocking constraint as posited by Aronoff (1976: 43) suggests that novel formations will not emerge if a synonymous word already exists in the language, since that would go against the principle of economy in language use. Thus, there is no such thing as **a stealer* because the word *thief* blocks its formation and spread. The blocking constraint is not imperative, but can be violated. Among the cases that come to mind are words such as *the disconnect*, *the invite*, *the pay*, which do exist despite their corresponding (near-)synonyms *the disconnection*, *the invitation*, and *the payment*. For these examples numerous records are easily available in corpora of the English language. However, formations such as **the receive* or **the create* are perceived as ungrammatical and should theoretically be blocked by *the reception* and *the creation*. Nonetheless, there are numerous converted forms like these in Asian Englishes. It is the aim of this study to scrutinize these supposedly constrained formations.

1.1.2 A usage-based account of the deverbal converted noun construction

This work is situated within the usage-based approach to language. The key assumption of the usage-based approach is that a speaker's grammar is the result of their experience with language. As Bybee (2006: 711) states:

A usage-based view takes grammar to be the cognitive organization of one's experience with language. Aspects of that experience, for instance, the frequency of use of certain constructions or particular instances of constructions, have an impact on representation that is evidenced in speaker knowledge of conventionalized phrases and in language variation and change.

In her seminal work on the usage-based approach, Bybee (2010: 9) describes language as the result of the interaction of various domain-general processes. What this means is that language is simply another of the cognitive processes that a human being is capable of and that language is not stored or processed any differently from other knowledge humans might have. Consequently, our experience with language is subject to the same processing mechanisms that any other experience might be. The result of a speaker's experience with language is their grammar, which, in Diessel's (2007: 830) words,

is seen as an emergent system consisting of fluid categories and dynamic constraints that are in principle always changing under the influence of general cognitive and communicative pressures of language use.

What the notions of *emergent* and *dynamic* imply is that grammar depends on and can change with experience. Accordingly, grammar is susceptible to new information and individual speakers' grammars also differ because of their unique experiences with language.

One key factor in shaping grammar is thus the frequency of occurrence of linguistic phenomena. Phenomena with a high frequency of occurrence will generate large amounts of experience and hence be well represented in a speaker's grammar. Phenomena of low frequency of occurrence are generally assumed to be less well represented due to the lack of experience with them. Frequency is therefore a crucial determinant of language representation; it affects "the comprehension, production, and emergence of linguistic categories and rules" (ibid.: 109).

Repeated exposure to the same form—regardless of the length or abstractness of that particular form, be it morpheme or chunk or syntactic pattern—will lead to this form being strongly represented in the mind. The process of habit-formation or routinization that comes with such repeated exposure is called entrenchment. According to Blumenthal-Dramé (2012: 4), "entrenchment denotes the strength or autonomy of representation of a form-meaning pairing at a given level of abstraction in the cognitive system". The notion of entrenchment is, as Blumenthal-Dramé (ibid.: 1) says, "as powerful as it is problematic", mainly because there is no clear-cut definition of the concept. However, for the purposes of this study, the fairly broad definition given here will suffice. (For a detailed account of entrenchment, the reader is referred to Blumenthal-Dramé 2012.)

According to a usage-based account of language, grammar thus emerges as follows: Through the highly frequent usage of a pattern in language, speakers (and hearers, for that matter) will have more experience with that pattern. As a consequence, the pattern will be very familiar to them and it will be stored (better), i.e. (more deeply) entrenched, in their minds. That way, they will be able to access it (more) quickly during language perception and production. Through increased usage and a higher level of entrenchment resulting from the former abstractions and generalizations can be made. These abstractions result in constructions, the integral parts of grammar.

In the following, verb-to-noun conversion is understood as a construction in terms of Cognitive Construction Grammar, as proposed by Goldberg (1995). According to her account,

1 Introduction

C is a CONSTRUCTION iff_{def} C is a form-meaning pair $\langle F_i, S_i \rangle$ such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other previously established constructions.

Additionally, constructions can also be viewed as such—even if they are completely predictable—if they “occur with sufficient frequency” (Goldberg 2006: 5). This modification of the original definition helps to accommodate findings that highly frequent items of language are more easily accessible than less frequent ones (cf. e.g. Arnon and Snider 2010).²

More precisely, conversion is interpreted as the embedding of an atomic and substantive construction, a verb, in an atomic and schematic construction, the nominal frame, as shown in 1.2.

The DEVERBAL CONVERTED NOUN construction: $[V]_N$ (1.2)

Formally, the result looks like an atomic and substantive construction. Nonetheless, the syntactic context reveals that in actual fact the verb has been inserted in the nominal frame. A fundamental mechanism in decoding the meaning of the DEVERBAL CONVERTED NOUN construction is coercion, the reinterpretation of the meaning side of a construction to fit the form side of it (cf. Lauwers 2008: 166). Coercion is facilitated by the fact that “[i]nterpretation favors syntactic meaning over lexical meaning” (Michaelis 2004: 62), which means that the nominal context into which a verb is inserted wins out over the verbal meaning that the verb would have in isolation.

Pinning down the meaning of the DEVERBAL NOUN construction is challenging, because, “[i]n contrast to typical word-formation patterns, the concept type [of a converted form] is not overtly marked and is therefore less prominent” (Schmid 2011: 194). Usually, profiling, i.e. highlighting what is important about a concept, is achieved by means of morphological material, yet, since conversion does without morphological material, profiling is less overt (cf. *ibid.*: 194–195). The main difference between verbs and nouns is taken to be one of reification (cf. Langacker 1987).

1.1.3 Varieties of Englishes

Situating the varieties in the Dynamic Model

The classification of varieties of English has for a long time followed Kachru's (1985) model of Three Circles, the inner, the outer, and the expanding circle. “While this classification

²However, the question of what constitutes “sufficient frequency” remains unclear. For a critical review of the frequency argument cf. Fahrner (2016).

was very useful when it first appeared, historical events have overtaken it, not least in the Southeast Asian region”, with which this study is concerned (Kirkpatrick 2012: 16). Current models of varieties and variety genesis are more process-oriented. The most prominent of these models is the Dynamic Model as proposed by Schneider (2003, 2007).³ Despite the observable differences between the New English varieties, his model of language evolution is based on the idea of an underlying process common to all emergent varieties of English. The main assumption of the Dynamic Model is that “the emergence of PCEs [Post-colonial Englishes] is an identity-driven process of linguistic convergence” (Schneider 2007: 30) which manifests itself in a “sequence of characteristic stages of identity rewritings and associated linguistic changes” (ibid.: 29) through which every variety progresses.

The stages of the Dynamic Model describe how groups of settlers and the indigenous people of the regions in question gradually converge, not only politically and socially, but also linguistically. This process of accommodation results in the genesis of New English varieties and is guided by extralinguistic factors (such as political developments), characteristic identity constructions of the settler and indigenous groups, and the “sociolinguistic determinants of the contact setting” (ibid.: 31). The five stages of the model are briefly described in the following.

Stage 1 In the **foundation** phase, English is brought to a new territory by a small group of settlers who in all respects associate strongly with their mother country. Contact between indigenous groups and settlers is for “exclusively utilitarian purposes” (ibid.: 34) so that language contact is minimal.

Stage 2 During the stage of **exonormative stabilization**, language contact becomes more frequent as English is used in more and more domains (administration, law, education etc.). Both indigenous and settler groups experience the contact with the other group as enriching, which gives rise to increasing numbers of bilingual speakers among the indigenous population. It is in this phase that “earliest structural features typical of local usage emerge” (ibid.: 40).

Stage 3 The stage of **nativization** is the most important and central phase. At this stage colonies usually gain independence, both politically as well as linguistically. Regardless of their origin, all residents are united by the fact that they inhabit the same territory. There is regular contact between all groups and this in turn promotes the

³For a detailed critique of various classifications of varieties of English, see for example Buschfeld (2013: 43–49, 190–198, 2014: 189–198).

1 Introduction

emergence of a new variety. Characteristics of this phase are a “marked local accent”, “new wordformation products”, “alternative morphosyntactic behavior” (Schneider 2007: 44–48) and the like. An example of a variety at this stage is Hong Kong English (cf. *ibid.*: 133–139). Although it still exhibits features characteristic of stage 2, it has entered stage 3 already in the 1960s, according to Schneider (*ibid.*: 133, 135). (See below for a more detailed description of HKE.) Hong Kong Island was colonized in 1841–2 and English was mainly spread through the work of missionaries. In 1898, Hong Kong was leased to the British for 99 years, thus facilitating the dominance of British English. With the growing wealth and internationalization in the second half of the 20th century came an increase in the proficiency of English which resulted in the emergence of a new variety. This variety is marked by lexical borrowings from Chinese (due to Cantonese immigration in the 20th century), new compounds, semantic shifts, distinctive syntactic features and a characteristic accent. Although Hong Kong is not under British rule any more, English is still a co-official language.

Stage 4 The main characteristic of the stage of **endonormative stabilization** is the increasing self-reliance of the former colony, particularly as regards language. Where “full integration” is an important aim for society, the “gradual adoption and acceptance of local forms of English” is what is observable in the linguistic domain (*ibid.*: 49). At this stage, linguistic heterogeneity is often ignored and the emergent variety of English is codified so as to promote its homogeneity (cf. *ibid.*: 51). A variety that finds itself at this stage is Singapore English (cf. *ibid.*: 153–161). Singapore was an important outpost for the British East India Company and in 1826 became part of the Straits Settlement. In the following decades, Singapore not only became a Crown colony, but also saw an enormous influx of workers, mostly of Southern Chinese origin, as well as a drastic increase in its economic wealth. After the Japanese occupation during the Second World War and independence in 1965, Singapore experienced a phase of modernization and economic growth. At the same time, Singaporean politics advocated for its characteristic language policy which requires that every citizen know one ethnic language (Mandarin for people of Chinese descent, Tamil for Indians, Malay for Malays) as well as English. Particularly among younger people, English is now used in a broad range of domains, both formal and informal. This has brought about a distinct form of English which is characterized by new features in phonology, morphology, syntax, and new word formations as well as semantic shifts. (For a more detailed description of SgE see below).

Another stage 4 variety is Indian English (cf. Mukherjee and Gries 2009: 33; Schneider 2007: 161–173). The beginnings of English in India date back to 1600, when the East India Company started their trading activities in South Asia. In 1858, the British Crown took over the rule from the East India Company, leading to a further and more systematic spread of the English language. After the political independence of India in 1947, English—contrary to all expectations—flourished. This is most likely due to the multilingual and multiethnic character of India. English serves as a language for communication across different ethnic groups, it is an “interethnically neutral link language” (ibid.: 167). No indigenous language has been accepted for this purpose. Today, IndE is characterized by a very typical pronunciation as well as lexical and morphosyntactic innovations. (IndE is described below in more detail.)

Stage 5 Internal differentiation indicates that a variety has reached the last of the five stages. Smaller groups emerge within one “overarching national identity” (ibid.: 53). As regards language, dialects and sociolects originate.

The following section presents the key aspects of the three Asian varieties investigated and gives reasons for this particular choice of varieties.

Choice of varieties

The data for this investigation represent five varieties: British English (BrE), US American English (USE), Indian English (IndE), Hong Kong English (HKE) and Singapore English (SgE). The Asian varieties are introduced hereafter before detailing the reasons for this choice. Instead of focusing on the historical details of the respective regions/countries, the idea is that of delineating the importance of the English language in these areas. This involves not only a description of the domains where English is used and the functions it fulfills, but also a summary of people’s attitudes towards the English language. The map in figure 1.1 helps the reader to locate the countries in which the varieties are spoken.

All three Asian varieties have traditionally been classified as ESL varieties, that is, as belonging to the group of varieties that has emerged in postcolonial settings and in contact with various substrate languages (also called the Outer Circle, cf. Kachru 1985). Before the varieties are described in more detail, a word about the functions the English language can fulfill in ESL contexts is in order. Following Srivastava (1994), it is possible to distinguish four functions. The *auxiliary function* prevails if English is mainly used to acquire knowledge through studying books. In this scenario, English could be called a ‘library language’. The *supplementary function* is drawn on in those cases where English is required for restricted purposes.

1 Introduction

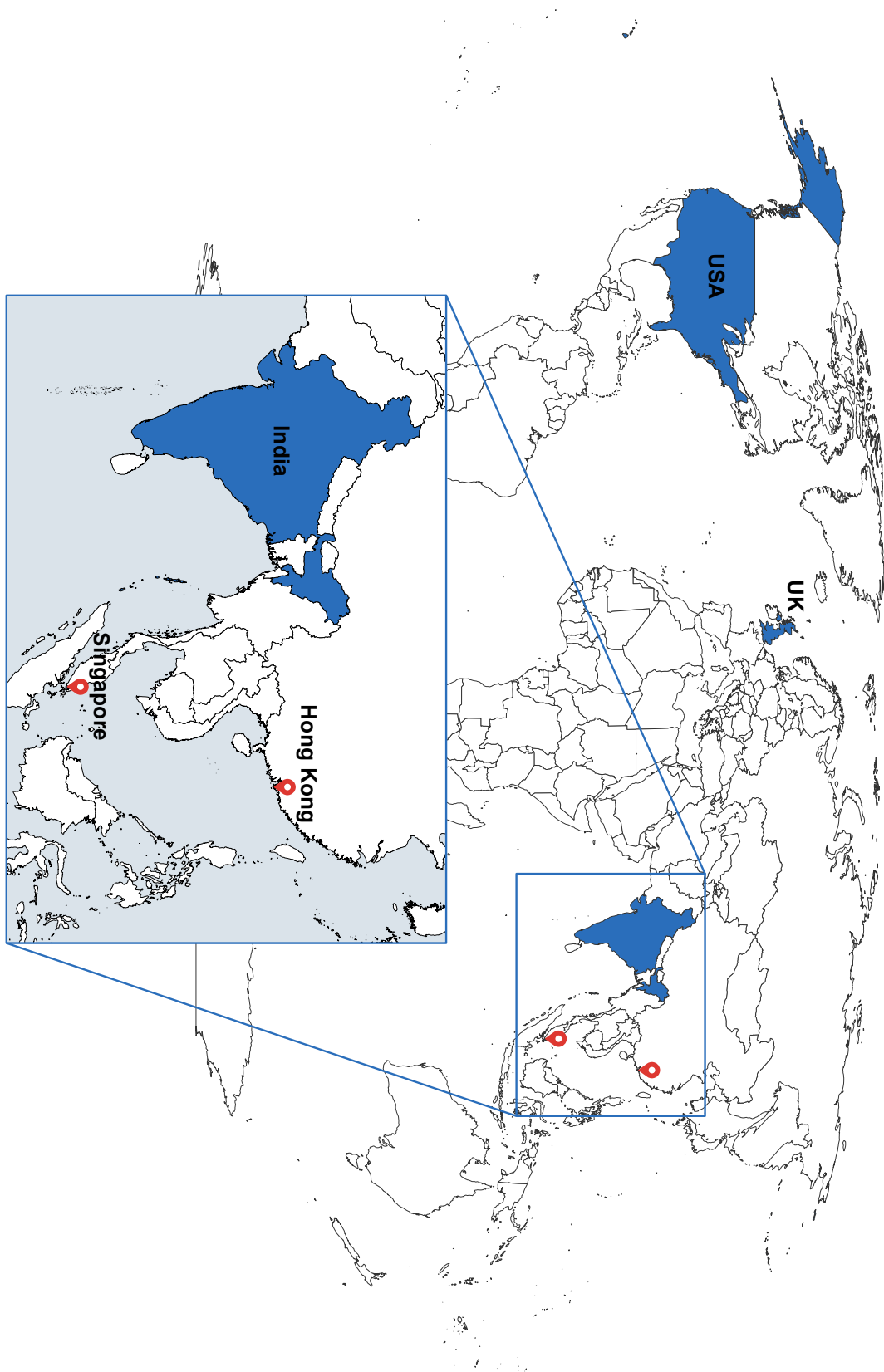


Figure 1.1: World map, with relevant countries highlighted, map by FreeVectorMaps.com (<http://freevectormaps.com>), adapted

Gargesh (2006: 92) mentions the example of taxi drivers in India. In these situations, English is mostly a ‘vehicular language’. In contexts in which English as a ‘link language’ is used in well-defined situations, complementing the L1, the *complementary function* of English is stressed. Finally, in those contexts where the L1 and English are used to the same degree, English assumes an *equative function*. In all three Asian contexts, English can be said to have been indigenized to at least such a degree that it fulfills the complementary function. For Singapore, it could be argued that English is also increasingly used in the equative function. For a “bird’s eye view” on the dynamics of English in Asia more generally, the reader is referred to Schneider (2014a).

Hong Kong English Hong Kong became a British colony after the First Opium War between China and Britain in 1841. In the Treaty of Nanking in 1842, Hong Kong was ceded to the British Crown. In 1860, after the Second Opium War, Kowloon and the New Territories were leased to the British. The lease expired in 1997 and Hong Kong was re-integrated into the People’s Republic of China as a Special Administrative Region (SAR). The designation as SAR stems from the fact that Hong Kong has kept a large amount of its colonial heritage, most importantly the capitalist economic system and its law system, modeled on the British one (cf. Bolton 2003: 50–51).

The spread of English in Hong Kong began only at a fairly late stage in the colonial history. Merely a small group of people, who Luke and Richards (1982: 51) call “linguistic middle men”, were fluent in both English and Cantonese. For the majority of residents, English “was not really in contact with the languages of the indigenous populations in domestic environments” (Gisborne 2009: 150). Up to the 1960s, an English-medium secondary education was “typically” restricted to the children of “only the socially privileged” (Bolton 2000: 269). These circumstances have often been referred to as ‘elitist bilingualism’ (cf. *ibid.*). Change came in 1974, when free, compulsory primary education in English was introduced (cf. Bolton 2012: 226). Four years later, free secondary education was established. This led to a rapid spread of English, giving rise to ‘mass bilingualism’ (cf. Bolton 2000: 269). In 1998, after the Handover (of Hong Kong to the Chinese), schools were obliged to revert back to Chinese-medium instruction. Subsequent continued protests brought change, with 114 schools teaching in English, the rest (approx. 300) remaining Chinese-medium schools (cf. Evans 2000: 185–186). A little over ten years later this policy has been relaxed and since the academic year 2010–11, schools can choose (“according to the needs and abilities of their students”) whether they prefer Chinese or English as the medium of instruction (Lai 2012: 85).

1 Introduction

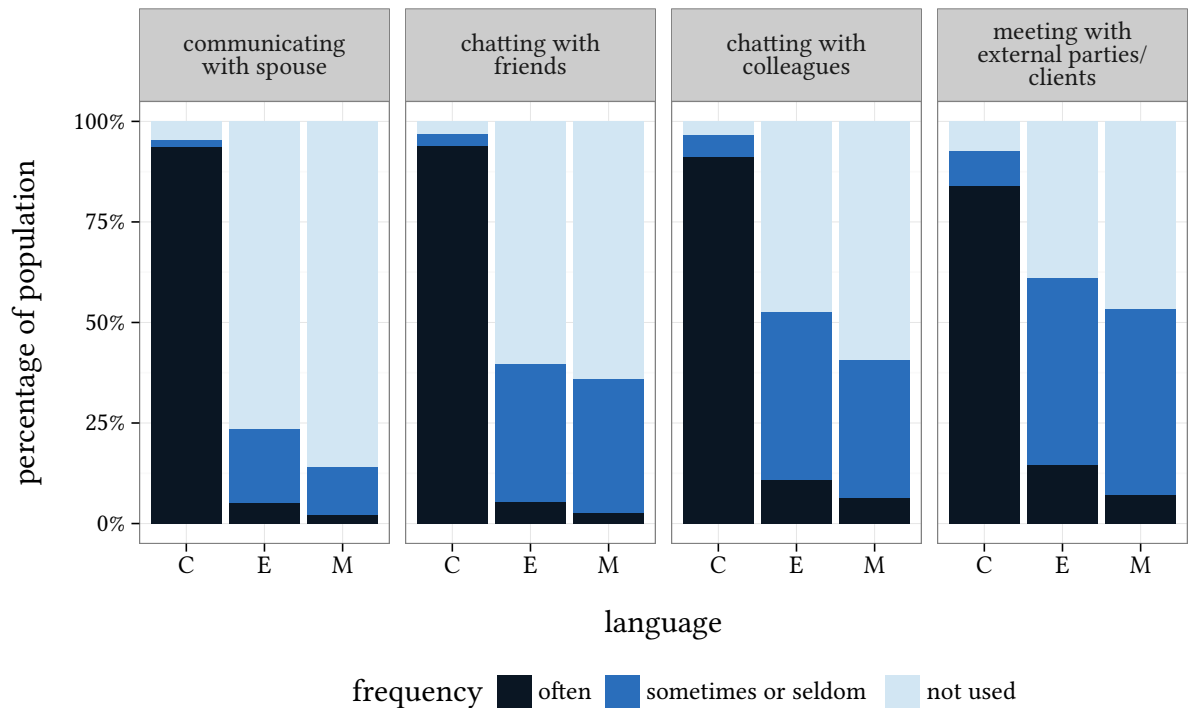


Figure 1.2: Languages spoken in different contexts in HK, data from Census and Statistics Department, Hong Kong Special Administrative Region (2013: 19–24). C=Cantonese, E=English, M=Mandarin

Today, English is still an official language in Hong Kong, next to Cantonese and Mandarin (also called Putonghua) according to Article 9 of *The Basic Law of the Hong Kong Special Administrative Region of the People’s Republic of China*. The official language policy since 1995 is one of trilingualism (fluency in Cantonese, English, and Putonghua) and biliteracy (Chinese and English). English is mainly used in administration, the legal system, business contexts, and higher education (cf. Evans 2010: 165). In spoken interaction, Cantonese prevails. Figure 1.2 illustrates the use of the three languages in various oral communicative situations. As is apparent from this chart, “[t]he sociolinguistic situation in Hong Kong is increasingly triglossic (in terms of Cantonese, Putonghua, and English), each language serving distinct functions” (Pang 2003: 17, also cf. Evans 2010: 160).⁴ This ‘division of labour’ can be traced back to the circumstance that “most Hong Kong residents have an emotional attachment to Cantonese and perceive English and Mandarin to be languages which have instrumental

⁴The growing triglossia has to be taken with a grain of salt considering that even in a very formal domain such as business meetings over 75% of the population still use Cantonese “often” and at the same time over 35% do not use English in this situation.

value, but which they are not particularly attached to” (Gisborne 2009: 152–153; also cf. Lai 2005). This is reflected in the increased usage of English and Mandarin in the business context, and is due to the fact that from very early on in the colonial history, “English in Hong Kong has first been the language of the governing race, and therefore of law and administration, then the language of international trade and finance”, but never a language largely used in private communicative settings, e.g. between friends or family members (Pang 2003: 15).

English thus can be said to occupy an important function in terms of public domains of life (business, education etc.), whereas in the more private domains of life (family and friends) preference is given to Cantonese. This is only possible because the population of Hong Kong is linguistically speaking very homogeneous (unlike the population of Singapore, where English serves as an interethnic lingua franca), as the figures for native languages from the 2013 census show (cf. figure 1.4 on page 15). More than 90% of all Hong Kongers report to have Cantonese as their first language. Other Chinese dialects such as Hokkien and Teochew (cf. Gisborne 2009: 150) and also Mandarin are marginalized. It has to be noted, however, that Mandarin/Putonghua is becoming increasingly important in Hong Kong due to “Hong Kong’s reliance on the mainland”, mostly as regards economy (Lai 2012: 86, 101, 104–106). In a comparison of teenage students’ (15–17 years) attitudes towards the various languages in 2001 (cf. Lai 2005) and 2009, Lai (2012: 91) finds that “the overall attitude pattern toward the three spoken languages [was] the same” but that “attitudes toward Putonghua [were] significantly more positive in 2009 than 2001”. The latter involves both the integrative⁵ and the instrumental⁶ domain (cf. *ibid.*: 92). This is also visible in the census data (cf. Census and Statistics Department, Hong Kong Special Administrative Region 2013: 23), which show that Mandarin is used at least sometimes in business contexts by around 40% of the population (cf. figure 1.2).

The instrumental function of both English and Mandarin in HK is also evident from the discrepancy between the number of people who report English or Mandarin as their native languages (cf. figure 1.4) and the number of people who claim to have a solid (i.e. very good, good, or average) knowledge of these languages (cf. figure 1.3). On average, only 1.4% of Hong Kongers have English as their native language, but 60.6% judge their command of English to be at least of average. The same is true for Mandarin, which 3.2% of the population report to have as a first language. Nonetheless, 63.9% indicate that they

⁵The integrative orientation was measured by statements such as “I like Putonghua.”, “As a Hongkonger, I should be able to speak fluent Putonghua.”, or “A person who speaks fluent Putonghua is usually educated, intelligent and well-off.” (cf. Lai 2012: 94).

⁶The instrumental domain was tested by statements like “English will help me much in getting better career opportunities in the 21st Century.”

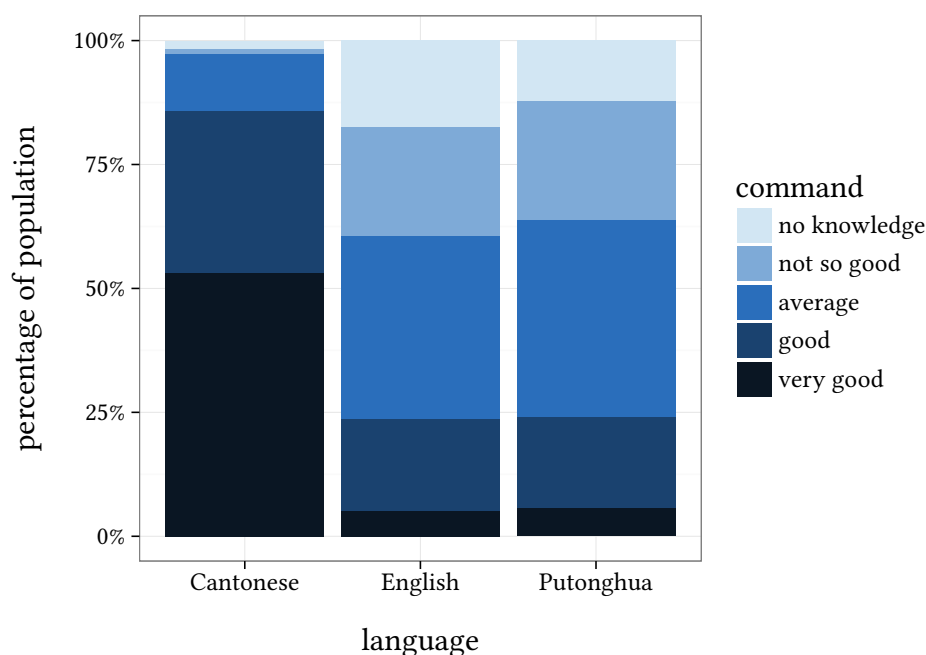


Figure 1.3: Command of spoken languages in Hong Kong, data from Census and Statistics Department, Hong Kong Special Administrative Region (2013: 14–16)

have at least an average command of it. As far as English is concerned, the high number of competent speakers is mainly attributable to the fact that English is indispensable in business contexts, as studies such as Evans’s (2010) and Chan’s (2013) show.

Notwithstanding the high number of people claiming to have a solid knowledge of English, figure 1.4 shows that English is not really gaining ground as a native language, contrary to what is observed in Singapore, for example. In Singapore, a language shift towards English is apparent (cf. figure 1.5 on page 17), in Hong Kong, however, the situation remains stable. What is similar to Singapore is the decreasing importance of other Chinese dialects such as Hokkien or Teochew (cf. Gisborne 2009: 150), with the exception of Putonghua, of course.

The fact that both English and Mandarin are not really native languages but rather instrumental languages leads to a strong orientation towards “outside standards” (Pang 2003: 17). In the case of English, this exonormative orientation is targeted towards the British English standard, as Lai (2012: 99) notes. Due to this strong orientation towards the British norm, the status of Hong Kong English as a variety of English in its own right has repeatedly been discussed. While Luke and Richards (1982: 55) and Johnson (1994: 182) in studies from the 80s and 90s oppose the idea of Hong Kong English—Luke and Richards (1982: 58) call

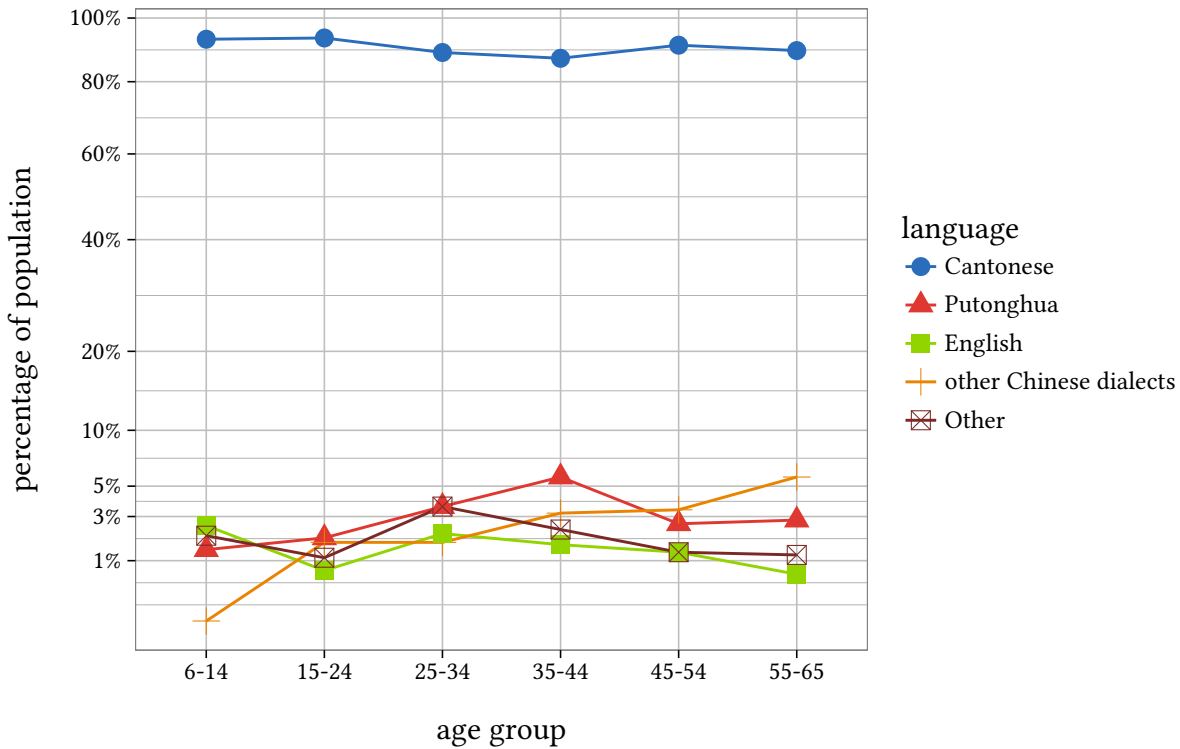


Figure 1.4: Native languages by age group in Hong Kong, data from Census and Statistics Department, Hong Kong Special Administrative Region (2013: 13). For better legibility, the y-axis is scaled.

it “an *auxiliary language*, rather than a second or foreign language”—newer studies such as Li (2000), Bolton (2003), Schneider (2007), Evans (2009), and Groves (2011) agree that Hong Kong English is becoming ever more nativized and thus deserves to be called a variety of English (with the exception of Pang 2003, who argues that HKE is not indigenized yet). Not only has HKE developed its own characteristic features, but there is also a strong ‘complaint tradition’ with standards reported to be declining (cf. Evans 2010: 162), which according to Schneider (2007: 56) is an indicator for a variety in the nativization phase of the Dynamic Model.

As far as phonological, grammatical and syntactic features of Hong Kong English are concerned, the reader is referred to Bolton (2003), Bolton (2002), Gisborne (2009), Hung (2012), and Setter et al. (2010) for comprehensive descriptions. The development of the variety is traced in great detail by Evans (2009, 2014, 2015a). More on language policy can be found in Bolton (2012) and Evans (2013).

Singapore English Singapore became a British settlement in 1819, after Sir Thomas Stamford Raffles established a dependency of the East India Company on the island. Its location in the straight of Malacca was of high strategic importance to the East India Company (cf. Deterding 2007: 2). In 1867, Singapore became a Crown Colony (cf. Leimgruber 2013c: 3). During this period, Singaporeans spoke either Bazaar Malay (a Malay pidgin) or “a simplified form of Hokkien” (Bolton and Ng 2014: 309). In 1963, Singapore gained independence from Britain and fused with Malaysia (cf. Deterding 2007: 2). This union was short-lived and in 1965 Singapore became a sovereign state. After the independence from Malaysia, in order to become an international competitor and also to unite the different ethnicities present in the country, it was decided that the national language should be English (cf. Wee 2013: 105–107).

Since then, English has served as an interethnic lingua franca and language policy has always been designed so as to artificially keep the status of English as the lingua franca upright, encouraging that children learn English in addition to an ethnic mother tongue. Among the multiethnic Singaporean population, English has to “remain ethnically neutral” and serve as the “non-Asian ‘other’” (Lim et al. 2010: 5–6). After the independence, after a brief period in which English or one of the official mother tongues (see below) served as media of instruction in schools, most non-English medium instruction schools “were closed because of falling student numbers” (Bolton and Ng 2014: 309). By 1987, English had thus become the main medium of instruction (cf. *ibid.*). At the same time that English became more dominant, the emergence of Singlish, the basilectal variant of SgE, was first noted (cf. Platt et al. 1983). In order to discourage the population from speaking the basilectal variant of English, in 2000 the *Speak Good English Movement* (SGEM) was launched. “Official motivations” for the campaign are “concerns about academic achievement, economic advancement, intelligibility, [and] Singapore’s national image” (Bolton and Ng 2014: 315). The launching of a campaign the aim of which is to preserve high standards of English aptly illustrates the importance of the English language for the country and its citizens, who understand English to be the “language of socio-economic mobility” (Lim et al. 2010: 5–6).

Next to English, the lingua franca, there are three official mother tongues of equal constitutional status (cf. Wee 2013: 107–108). These are assigned to the three major ethnic groups: “Mandarin for the Chinese, Malay for the Malays, and Tamil for the Indians” (Leimgruber 2013c: 12). The language policy in Singapore is summarized by Wee (2013: 109) as follows:

- i. “Recognizing a total of four official languages. Of the four, English is not given a status as a mother tongue.
- ii. Encouraging bilingualism in English and an ethnic mother tongue.
- iii. According a specific mother tongue to each of the major ethnic groups.”

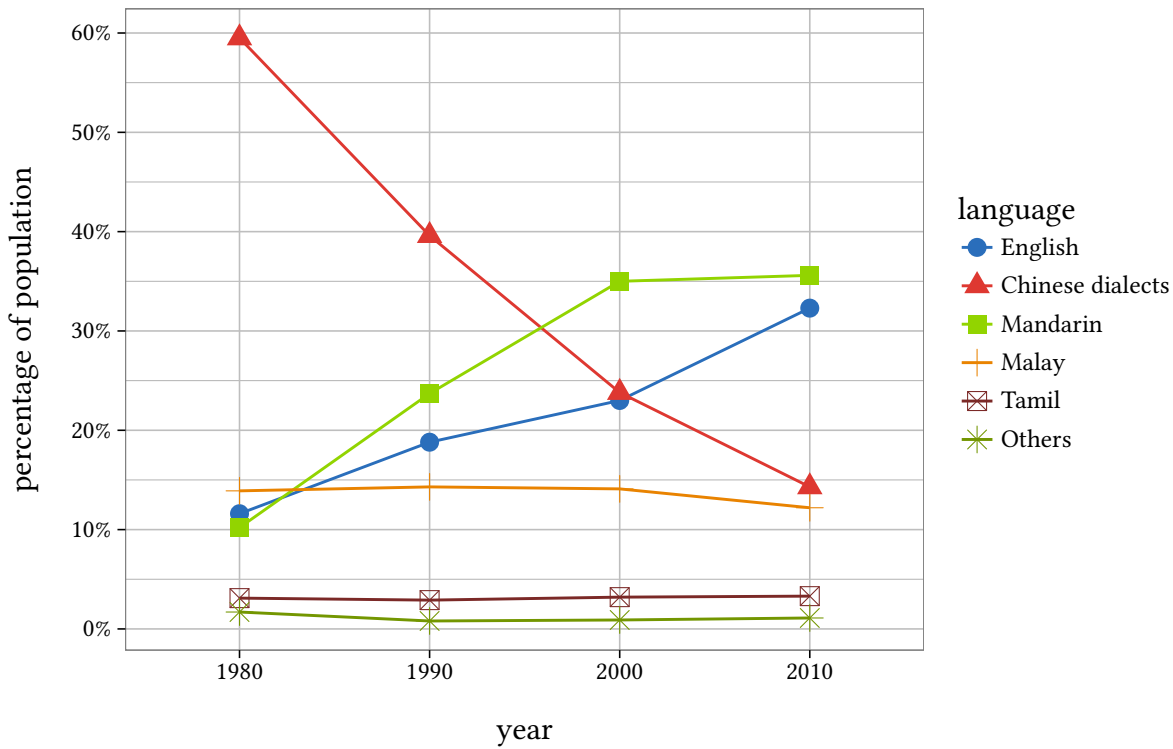


Figure 1.5: Languages most frequently spoken at home in Singapore, data from Leimgruber (2013a: 7)

Nevertheless, these official languages were and still are not necessarily the inhabitants' mother tongues, particularly as far as the Chinese group is concerned. At the time of independence, a considerable part of the Chinese group had Southern Chinese dialects such as Hokkien, Teochew, or Cantonese as their mother tongues (cf. Bolton and Ng 2014: 308–309).⁷ Nonetheless, Mandarin was chosen as a 'mother tongue' for the ethnically Chinese because of its important political function in "unifying" the different groups of Chinese (Goh 2013: 127). Since being accorded the status of an official mother tongue and the launch of the *Speak Mandarin Campaign* (SMC) in 1979, Mandarin has displaced other Chinese dialects, as is visible in figure 1.5 (cf. Bolton and Ng 2014: 311; Leimgruber 2013b: 232–233). Therefore an age-related difference as far as proficiency in Mandarin is concerned can be observed (cf. Goh 2013: 133), with younger people speaking more Mandarin than older people.

It can hence be concluded that "the influence of Mandarin in S[g]E has to be relatively recent, though no doubt increasingly significant" (Ansaldo 2004: 135). This is all the more likely considering that 74% of all Singaporeans are ethnically Chinese (13.4% are Malays,

⁷The same discrepancy can be observed for the other groups. Until this day, Indians also use Sanskrit and Malays also make use of Arabic, mostly for religious and cultural purposes (cf. Vaish 2008: 457–462).

1 Introduction

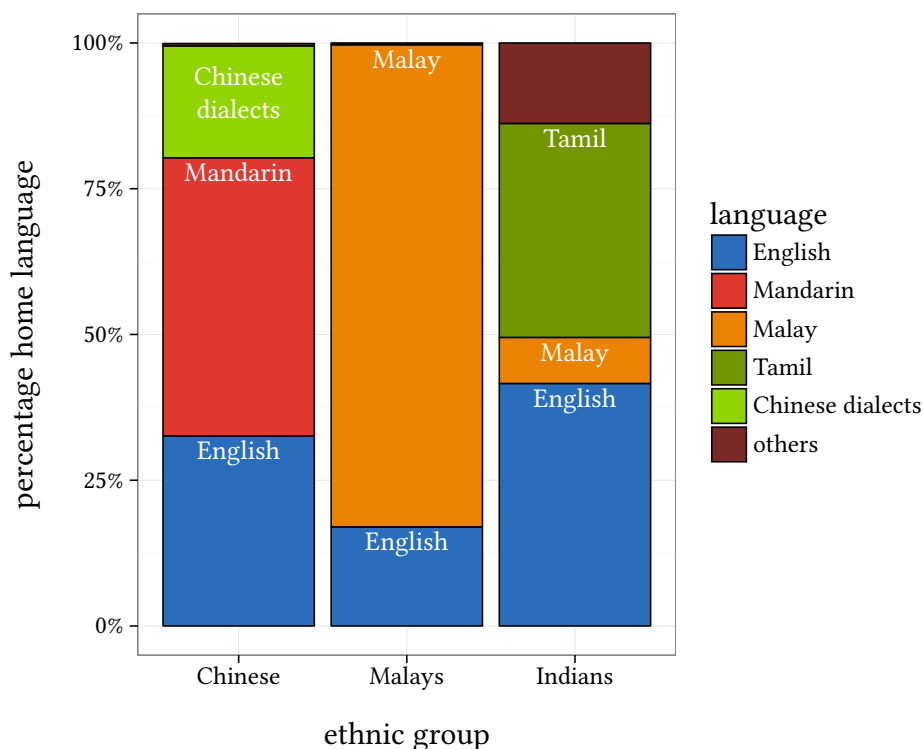


Figure 1.6: Languages most frequently spoken at home in Singapore by ethnicity, data from Wong (2011)

9.2% are Indians, cf. Wong 2011). The dominance of Chinese in Singapore is already visible in Platt et al.'s (1983: 44–45) study that found that a Malay Singaporean in the study had already adopted Chinese features as early as 1983 as a result of the dominance of Chinese. Today, the high instrumental value of English and Mandarin and the “personal gain and social mobility” that result from knowing both are strong motivators for the shift towards these two languages (Bolton and Ng 2014: 315). As Siemund et al. (2014: 350) in a study on the languages used by university and polytechnic students in Singapore show, bilingualism is the norm among students (born between 1984 and 1998), and the languages most frequently spoken by students are English (all participants) and Mandarin (270 out of 300 participants). The most frequent language combinations are English/Mandarin, English/Hokkien/Mandarin, English/Cantonese/Mandarin, and English/Malay (cf. *ibid.*: 351). All these findings are indicative of the language shift towards English and Mandarin that is taking place, particularly when keeping in mind that “today’s students form tomorrow’s high-income groups who are likely to be the social trendsetters” (*ibid.*: 360).

That the current language policy of bilingualism in English and a mother tongue does not seem to reflect actual language use and proficiency can also be gleaned from Tan's (2014) study. Tan (*ibid.*) in a questionnaire study with 436 Singaporean participants of all three ethnic groups demonstrates the increasing importance of English in basically every domain of life. The results suggest that particularly for the youngest age group, English "has overtaken [all] other languages" (*ibid.*: 334) as a means of expressing and identifying oneself and communicating with friends and family. The fact that "[c]lose to 70 per cent of the young Chinese participants" and even higher numbers of Malay (74%) and Indian (100%) participants "want to use English to communicate with their children" (*ibid.*: 330–331) indicates that she is right when she claims that "English can and should be thought of as a mother tongue for Singaporeans" (*ibid.*: 319), despite official policy, at least in the years to come. As figure 1.6 shows, this language shift towards English affects all ethnic groups, although the Malay part of the population seems to be shifting more slowly than the groups of Chinese and Indians.

The language shift towards English is particularly likely to happen considering the dominance that English has even in the most private domains such as communicating with friends and partners (*cf. ibid.*: 334). The ethnic mother tongues are predominantly reserved for religious purposes and to communicate with family and close friends, as Vaish's (2008) study of 10-year-olds' language use shows. What is remarkable about Vaish's (*ibid.*: 458) work is the finding that children use English more often for silent prayer than for praying at the church or mosque or temple, which shows that particularly the younger Singaporeans have already largely shifted to English as their dominant language. This situation contrasts strongly with what has been shown above for Hong Kong, where the majority of the population prefers Cantonese to talk to partners and friends (*cf. figure 1.2*).

The question whether Singapore English is a variety in its own right is hence undisputed. It has long entered the stage of endonormative stabilization. A basilectal form of English, Singlish, emerged as early as in the 1980s. Studies dating from this period already document Singlish extensively (*cf. Platt et al. 1983*). Although the government is trying to counteract the use of Singlish with the *Speak Good English Movement* initiated in 2000 (*cf. Low 2012: 26*), Siemund et al. (2014: 341), among others, claim that "it really is a distinct variety with special phonology, morphosyntax, and vocabulary". In the same vein Wee (2013: 114) asserts that "in actual fact, the emergence of Singlish is an indicator for the successful nativization of English in that territory". One could even go so far as to claim that the existence of the basilectal variant, Singlish, alongside the acrolectal standard(izing) SgE is an example of differentiation that is typical of the final phase of the Dynamic Model (*cf. Kirkpatrick 2012: 17; Wee 2014*).

For a detailed discussion of the features of Singapore English the reader is referred to a recent special issue of the journal *World Englishes* (Vol. 33, No. 3, 2014), Deterding (2007), Leimgruber (2013a), Lim (2004), Low (2012), and Schneider (2007: 153–161). Bao (2005) and Gut (2009) provide studies of verb morphology and the aspectual system, respectively. More on language development can be found in Alsagoff (2012), Alsagoff (2010), and Ansaldo (2004). Language policy is extensively covered in e.g. Leimgruber (2013b).

Indian English On December 31, 1600, Queen Elizabeth granted a charter to several merchants to trade with India. The charter led to the foundation of the East India Company, and in 1612, the first trading post was set up in Surat in India (cf. Lange 2012: 21). Over the course of that century, other posts were founded and conquered, such as Madras (1639), Bombay (1668) or Calcutta (1690). Between 1757 and 1857, the East India Company conquered large parts of India, eventually coming to control almost the entire sub-continent (including what are nowadays Pakistan and Bangladesh, cf. Sedlatschek 2009: 8–11). At roughly the same time, between 1780 and 1830, several missionary schools and colleges were founded, which led to an influx of British settlers and an increasing demand for English-speaking individuals (cf. *ibid.*: 11). In 1858, after the Great Rebellion of 1857 against the East India Company rule, the British Crown seized control over India, which resulted in an even greater importance of the English language in India (cf. *ibid.*: 14–15). Colonial rule lasted until 1947, when India gained independence from the British Empire.

The English language was used in missionary schools already in the early 18th century. In 1835, after Thomas Macauley's *Minute of Indian Education*, written in the context of the debate over "the appropriate role of English" (*ibid.*: 12), English became the medium of instruction in secondary schools and in universities (first universities founded in 1857). In 1882, more than 60% of all primary schools were English-medium (cf. Kachru 1994: 507–508).

In 1947, after independence, Hindi was declared the official language (cf. Sedlatschek 2009: 17). Yet, English was too deeply rooted in India to be replaced immediately. English was therefore retained as official language, if only for a trial period of 15 years at first (starting in 1950 when the constitution was passed). In 1963, however, it was decided that English should remain, at first as a "co-official language" (1963–1967). Later English was declared an "associate official language" in the Official Languages Act of 1967 (cf. *ibid.*: 18–19).⁸ Nowadays, English is used in a wide variety of domains (see below), most notably in legislation and in the judicial system, where English is used exclusively (cf. Sailaja 2009: 5).

⁸There are several states in which English is the official language, for a detailed list cf. Sedlatschek (2009: 19–20).

The choice to retain English as the official language after independence was based on the fact that only English could assume the function of an interethnic lingua franca, or as Mukherjee (2007: 167) calls it, “a useful and inevitable pan-Indian link language”. Non-Hindi-speaking people (concentrated in the South of India) “thought that Hindi as an official language would offer unfair advantage to the people of the North and curtail their upward socio-economic mobility, and so they began to support the retention of English” (Gargesh 2006: 94).

Since then, languages have been categorized as so-called scheduled and non-scheduled languages. According to the *Eighth Schedule to the Constitution of India*, there is a total of 22 scheduled languages. These are complemented by 100 non-scheduled languages, English among them (cf. Census of India n.d.[a]). Many of the languages in India are Indo-European, but there are also languages from the Austro-Asiatic, Tibeto-Burman, and Dravidian language families (cf. figure 1.7; Census of India n.d.[d]).

It is not clear how many of India’s inhabitants are fluent in English; a very low estimate is at 5%. However, considering that the population of India is so large this figure would still make Indians the third largest group of English speakers behind the populations of the US and the UK (cf. Mukherjee 2007: 163). According to the representative India Human Development Survey of 2005 (Desai et al. 2010: 95), 5% of the male population aged 15 to 49 years and 3% of the female population of the same age are fluent in English and 28% and 17%, respectively, have “some” knowledge of English.⁹ If the average of male and female Indians with “some” knowledge of English (22.5%) is multiplied by the number of inhabitants (roughly 1.2 billion), one arrives at 270 million at least somewhat competent English speakers in India. Regardless of the exact numbers, competence in English is certainly increasing in India. In the 1971 census, approximately 192,000 people claimed to have English as their native language. In 2001, the figure has risen to 226,000 people (cf. Census of India n.d.[c]).

The ten languages most frequently reported as native languages in the 2001 census (cf. Census of India n.d.[b]) are shown in figure 1.7. The language policy since 1957 is one of trilingualism. According to the *Three Language Formula*, every Indian is expected to know three languages: first, a native language (a regional language), second, in Hindi-speaking states another modern Indian language and in non-Hindi-speaking states Hindi, and third, English (cf. Gargesh 2006: 94–95). However, as Sedlatschek (2009: 20) points out, “[t]here are marked differences in the ways that individual states have implemented the Three Language Formula”.

⁹The survey questions evaluated whether participants “speak no English” or “speak some English” or “converse fluently” (Desai et al. 2010: 85). These categories are not clear-cut, which makes them prone to subjective (and therefore potential mis-)interpretation by the respondent.

1 Introduction

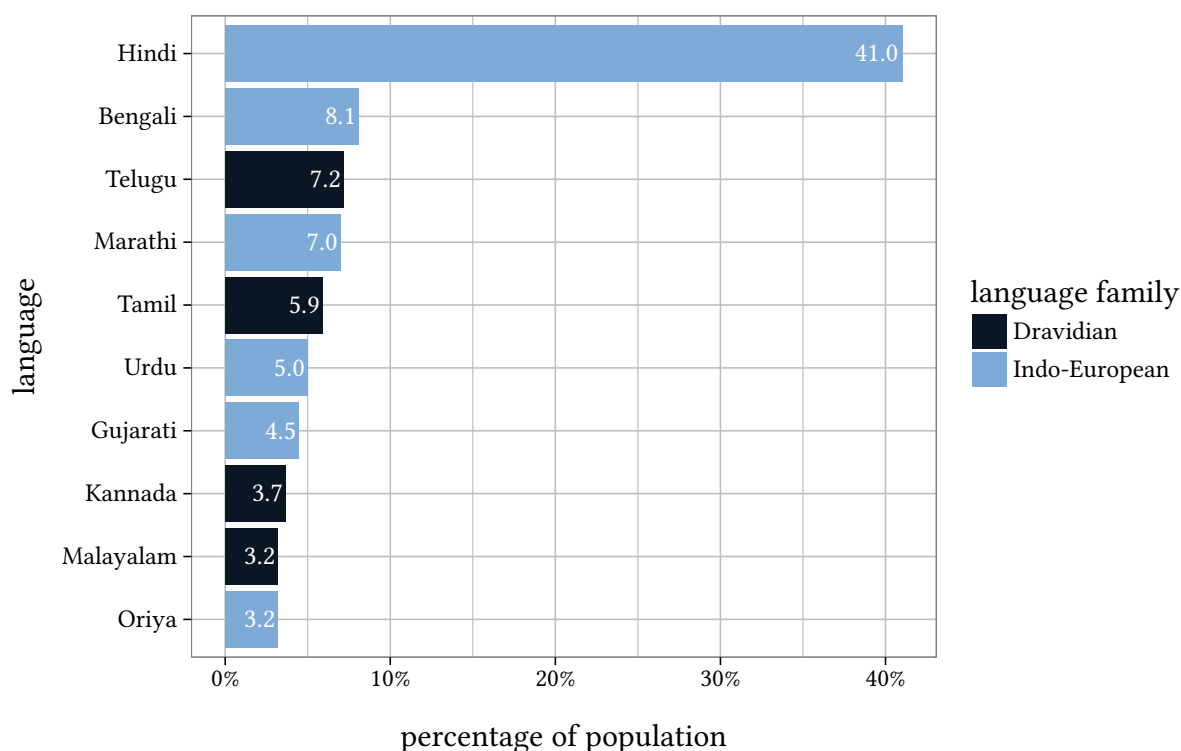


Figure 1.7: Ten most frequently spoken (scheduled) languages in India, data from Census of India (n.d.[b])

Even though many Indian linguists (when asked by fellow linguist David Crystal) think that “around a third of the population [are] these days capable of carrying on a domestic conversation in English” (Crystal 2008: 5), it can generally be assumed that the level of proficiency in English is strongly related to socioeconomic status as well as regional and educational background (cf. Desai et al. 2010: 95–96). The “affluent and influential sections of Indian society” are heavily associated with a higher language proficiency, which accounts for the high prestige that English enjoys (Sedlatschek 2009: 2). Figure 1.8 depicts the number of children enrolled in English-medium schools per state and reveals vast regional disparities. Yet, education in English also depends on other factors not visualized in the map: “English medium enrolment is the most prevalent in metropolitan areas (32 per cent), among families with a college graduate (32 per cent), and among the top income quintile (25 per cent)” (Desai et al. 2010: 86).

While English-medium instruction is not compulsory (cf. Sedlatschek 2009: 20), it is “overwhelmingly the desired medium of education” (at least in some parts of India), mostly due to the instrumental value that Indians generally attach to the English language (Gargesh

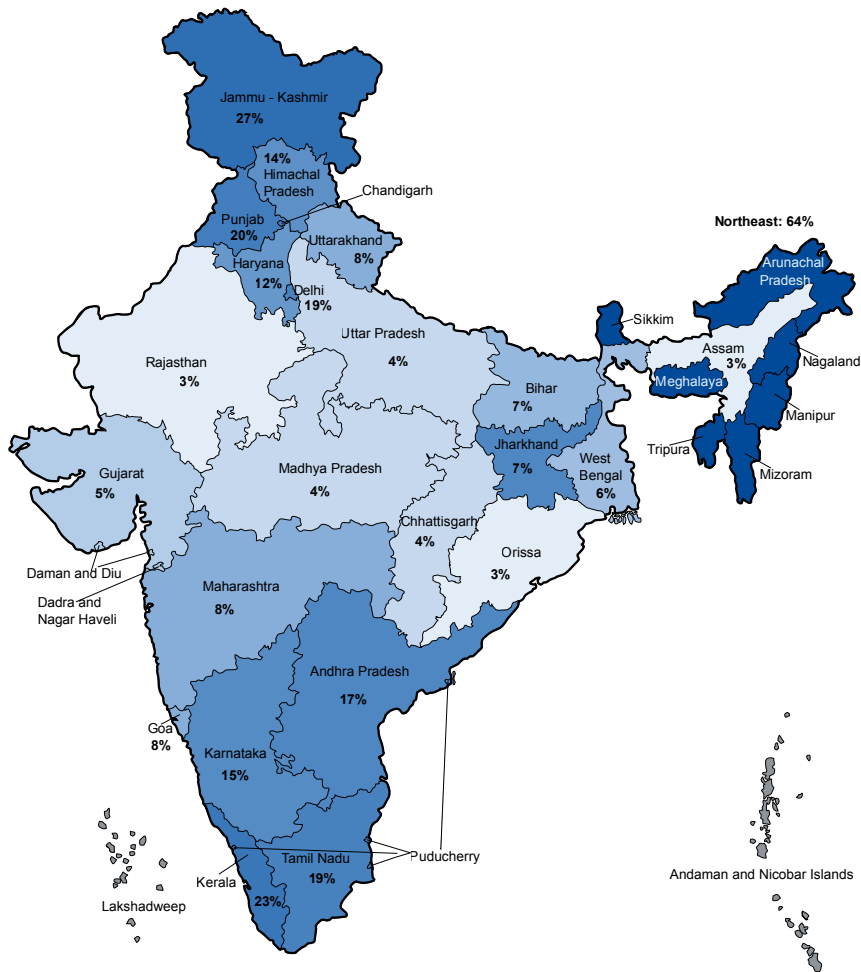


Figure 1.8: Percentage of children (6-14 years) enrolled in English-medium schools, data from Desai et al. (2010: 86), map by Rajeshodayanchal at Malayalam Wikipedia (2011), adapted

2006: 101). However, contrary to what is observed for Hong Kong and similarly to what has been shown for Singapore, English is not just perceived as a means to obtain a “competitive advantage in the global society” (Das 2002: 20). A summary of various studies concerned with language attitudes and language usage (cf. Sedlatschek 2009: 22–24) reveals that English is not only used in the public domain but also with family, friends and neighbors, that is, in more intimate contexts.

As far as the dynamics between the two main languages, English and Hindi, are concerned, English seems to be of greater relevance than Hindi, not only in the non-Hindi-speaking South, but also in the rest of India. Already in 1987, Sridhar (1988: 314–315) ob-

1 Introduction

serves that “the spread of Hindi [through the Three Language Formula] has not resulted in any appreciable replacement of English”. On the contrary, Sridhar (1988: 314–315) notes that

English continues to command prestige and to symbolize education, power, and modernity. Its international currency and association with the great developments in science and technology give it certain advantages that Hindi cannot match. Forty years of independence have not diminished the role of English; if anything, its importance in the life of the nation has grown substantially.

Even though almost thirty years old, it seems that this quote is all the more valid today.

From the functions that the English language performs in India, it can be concluded that Indian English is largely an endonormatively stabilized variety, even though some scholars, most notably Schneider (2007: 171), situate IndE at the nativization stage. However, unlike SgE, it still shows traces of the preceding phase (cf. Mukherjee 2007: 170). A further striking difference between IndE and SgE is the number of speakers claiming to have English as their native language. This number is much higher in SgE, indicating that the degree of indigenization of English must be higher in Singapore than in India, even though both varieties might be classified as endonormatively stabilized varieties.¹⁰

Among the works that provide comprehensive overviews of Indian English features are Sailaja (2009) and Sedlatschek (2009). Phonology is treated in Gargesh (2008), while syntax is the object of study in Lange (2012). Further morphosyntactic studies include Hoffmann et al. (2011) and Mukherjee (2009b) as well as Schilk (2011).

Overview: Asian Englishes

Table 1.1 offers a summary of the preceding sections describing the Asian varieties analyzed. It takes up the most relevant aspects, namely, the official languages, the current language policy, the main domains of use and functions of English, as well as the degree of institutionalization operationalized by the developmental stage.

Summarizing the above, this selection of varieties is anticipated to offer an intriguing picture of V>N conversion in World Englishes (cf. Mukherjee and Gries 2009: 31). From a methodological perspective, the availability of corpora specifically designed for the purpose of comparing New English varieties poses a big advantage. From a linguistic point of view, all three Asian varieties are based on British English.

Hong Kong English and Singapore English are representatives of Asian Englishes and show similar contact ecologies. Both varieties have emerged in contact with a range of highly

¹⁰The adequacy of terms such as *endonormatively stabilized variety* or *ESL variety* will be discussed at a later point, drawing on the results from the corpus and experimental study (cf. chapter 8 and section 9.4).

Table 1.1: Overview of the Asian Englishes analyzed

	Hong Kong	India	Singapore
Time of British occupation	1841–1997	1612–1947	1819–1963
Other official languages	Cantonese, Putonghua	Hindi, furthermore: 21 scheduled languages	Mandarin, Malay, Tamil
Language policy	from 1995 on: trilingualism (Cantonese, English, Putonghua) and biliteracy (Chinese and English)	from 1968 on: Three Language Formula (one native language, Hindi, and English in secondary schools)	English and one mother tongue
Main areas of use of English	administration, law, business, higher education	administration, law, science, technology, education, media, with friends	all domains
Functions English fulfills	mainly complementary function	mainly complementary function	complementary and equative function
Developmental stage in Dynamic Model	(2–)3	3–4	4

analytic dialects of Chinese and are today in intensive contact with the equally analytic Mandarin. Nonetheless, as far as the degree of institutionalization of English in the respective regions/countries is concerned, HKE finds itself at stage 3 of Schneider's model but SgE has already moved on to stage 4.

Indian English has been selected to serve as a basis of comparison. It is another Asian variety and it shares the parent variety with HKE and SgE. However, its contact ecology is markedly different as the main substratum of IndE, Hindi, and many other minor contact languages on the sub-continent, are synthetic. Furthermore, English in India is institutionalized to a degree that is largely comparable to SgE, at least both varieties are often located at the same stage in the Dynamic Model.

This constellation thus allows for a comparison of varieties at different stages but with the same type of substratum (HKE vs. SgE) and also for a comparison of varieties at the same stage but with typologically different substrata (SgE vs. IndE). It is the aim of this study to see how these different linguistic ecologies compare with regard to conversion.

The Asian varieties are subsequently compared to the two most important native varieties, British English and US American English. British English as the parent variety of all three new varieties can be considered the point of departure. Nonetheless, in the last few decades, with growing globalization, the influence of US English has increased heavily. Particularly in the domain of digital media and the internet US English "has [acquired] a global reach and the potential to affect all other (standard and non-standard) varieties of English" (Mair 2013a: 259). US English has therefore been called the "hub of the World System of Englishes" (ibid.: 261). Together, BrE and USE are what Collins and Yao (2013: 479) call the "two inner circle 'super-varieties'" that exert influence on all varieties of English. It thus makes sense to include both varieties in the study.¹¹

¹¹Not unlike many other studies, this study is conducted without thoroughly assessing the exact nature of the influence of the media on the varieties investigated. While this might be problematic, it has to be acknowledged that a comprehensive investigation of the influence of different media (TV, internet etc.) and genres (sitcoms, social networks etc.) is unfeasible in the present context. For an overview of the challenges related to and recent studies concerned with the influence of media particularly on language change, the reader is referred to a special issue of the *Journal of Sociolinguistics* on media influence (Vol. 18, No. 2, 2014), particularly Sayers (2014).

1.2 Research questions

The interplay of substrate influence and institutionalization

Even though the substrate language/s shape/s the contact variety of English to a considerable extent, dominantly contact-based approaches to New Englishes have proven problematic (cf. e.g. Gut 2007: 347; Kirkpatrick and Moody 2009: 270; Laporte 2012: 286). As these studies have shown, it is not only the substrate languages but also the degree of institutionalization of English in a particular English-speaking community and the norm-orientation of that particular community that determine the presence and frequency profile of specific features. Along the same lines, the wide-spread classification of varieties into ENL (English as a native language), ESL (English as a second language), and EFL (English as a foreign language, originally conceived by Strang 1970: 17–19 and later taken up by Quirk et al. 1972: 3–4), as well as Kachru’s (1985) corresponding classification of varieties into Inner, Outer, and Expanding Circle, respectively, have proven too coarse to accurately represent the linguistic reality in many English-speaking areas around the world (cf. e.g. Biewer 2011; Deshors 2014; Edwards and Laporte 2015; Gilquin 2015; Gilquin and Granger 2011; also cf. Gilquin 2015 for a commented list of previous studies). Particularly the notion of ESL is problematic in that it subsumes all New Englishes under one heading due to the fact that the notion focusses primarily on the *emergence* of these varieties in post-colonial settings. Nonetheless, as the above-mentioned studies have shown and as this investigation will corroborate, it is rather the *development*, that is, the process of institutionalization and indigenization of individual varieties of English, that counts. This development is operationalized by drawing on the Dynamic Model (cf. Schneider 2007). Yet, the Dynamic Model does not specify in how far substrate influence persists beyond the nativization phase (cf. *ibid.*: 51–52). Schneider himself (*ibid.*: 45) does acknowledge that, in the nativization phase, innovations may be the result of either “transfer phenomena” from the substrate or “innovations caused by second-language acquisition processes”; however, the origin of innovations “is not of primary importance in the long run” to him. As this analysis of V>N conversion will show, speakers’ reliance on their L1, i.e. the substrate, remains a source for structural innovations even in a variety as advanced as SgE.

Furthermore, as the analysis will show, the level of individual features of New English varieties is at times difficult to integrate with the phases proposed by Schneider (2007), as these are seemingly too coarse to accurately capture the development of individual structural innovations. As Edwards and Laporte (2015: 161–162) point out, empirical results such as their findings on the preposition *into* (as well as the findings on conversion presented in later

chapters) beg the question of in how far the advanced stages of endonormative stabilization and differentiation can be reconciled with a purportedly less advanced, exonormative orientation of some structural features. In Edwards and Laporte's (2015) study, for example, IndE shows a profile that is markedly different from SgE but similar to HKE and to Dutch learner English (cf. *ibid.*: 162), despite IndE and SgE both being described as stage 4 varieties.

This leads to one of the main points of criticism against the Dynamic Model: that the model implies linearity in variety development. Yet, as e.g. Buschfeld (2014) and Edwards (2016) argue for Namibian English and English in the Netherlands respectively, there are varieties of English whose developmental trajectories have not followed the path outlined by Schneider. Contrary to the varieties covered in Schneider (2007), these two cases lack a colonial background (Netherlands) or have a mixed colonial background in which the British element has not dominated (Namibia). Schneider (2014b: 9) himself admits that "despite some similarities [between Expanding Circle varieties] it [= the Dynamic Model] is not well suited to grasp the vibrant developments of the Expanding Circle". However, it is not only in Expanding Circle contexts where the development of varieties can deviate from the assumed linear pathway. Mesthrie and Bhatt (2008: 35), on the basis of Gut's (2004) description of Nigerian English phonology, hypothesize that "a territory could move from phase 3 to 5, bypassing phase 4. This would be a territory in which English became nativised and subsequently differentiated into sub-dialects, without there being a commonly accepted endonormative standard." This aspect and also the aforementioned points of criticism are addressed in the final discussion with a view to the findings on verb-to-noun conversion.

The present study thus aims to explain the frequency profile of conversion in contact varieties of English by integrating the two determining factors of transfer from the substrate language(s) and degree of institutionalization of English. Both mechanisms are assumed to interact in shaping New Englishes. Substrates will influence the productivity of V>N conversion; particularly the Chinese dialects are projected to foster the process to a considerable extent. However, the frequency of V>N conversion is also hypothesized to vary with the degree of indigenization as operationalized by the above-mentioned stages in the Dynamic Model (cf. Schneider 2007). Less verb-to-noun conversion is expected for more advanced varieties, first, because conversion as a morphologically simple process is prone to be adopted primarily by less proficient speakers, and, second, because transfer from the substratum is hypothesized to be restricted in more advanced, endonormatively stabilized varieties.

The interaction of substrate transfer and institutionalization is the focus of chapter 6 and of chapter 7.

The role of usage frequency in verb-to-noun conversion

While the preceding section has stressed identity constructions as proposed by Schneider (*ibid.*) and transfer from contact languages as main factors in variety genesis, there is at least one other factor that merits careful attention: Usage frequency is a key determinant of language acquisition, processing, and change, and has recently come into focus (for an overview cf. e.g. Bybee 2010; Diessel 2007). However, the question of whether and, if so, in how far the frequency of various linguistic items determines and constrains conversion has not been answered yet. This study seeks to fill this gap by providing a usage-based account of verb-to-noun conversion.

The first aspect under investigation is the blocking constraint. It is assumed that the relative usage frequency of a near-synonym to the usage frequency of a converted form will matter crucially: the more frequent the near-synonym, the less frequent the converted form. The success of conversion, that is, the spread of the forms resulting from V>N conversion, will depend to a large extent on the relative frequency with which the blocking lexeme occurs. That is, in contexts where the usage frequency of the near-synonym, the blocking lexeme, is relatively low, the blocking constraint can be overridden and the converted form can establish itself alongside the near-synonymous form. Hence, the lower the ratio of blocking lexeme to newly coined lexeme is, the better the chances for successful verb-to-noun conversion are.

Secondly, entrenchment as a direct function of frequency is predicted to influence the productivity of conversion (cf. section 2.1.2). A form is highly entrenched if it is stored in the brain in such a way that it is easily accessible and retrievable. High usage frequency leads to deeper entrenchment. Thus, the frequency with which a potentially converted form occurs in its original word class can be hypothesized to influence the productivity of conversion. Low-frequency forms will convert more easily because they are less entrenched, which means that they will not be associated with a certain word class as strongly as high-frequency forms. High-frequency forms can be expected to convert less easily considering that they are strongly associated with the base word class. The phenomenon that more frequent forms change less quickly, the so-called conserving effect of frequency (cf. Bybee 2010), has been observed for conversion by e.g. Teddiman (2012).

A further point that is directly related to frequency of occurrence and productivity is the notion of acceptability. The more often a form occurs within a speech community, the more likely it is that a speaker will accept this form as part of their language. When asked for the acceptability or grammaticality of an innovative form, a speaker's rating will largely depend on how familiar they are with the form.

Moreover, frequency of use also plays an important role in determining the spread of a converted form. Linguistic elements that co-occur frequently are generally not processed individually but as larger units, so-called chunks (cf. section 3.1.2). Embedding a novel conversion in a frequently used chunk will help spread the innovation.

These mechanisms are illustrated in chapters 5, 6, and 8.

Frequency from a cross-variety perspective

A further question that this study explores is whether the role of frequency in influencing the emergence and spread of converted forms as pointed out above plays out similarly in native and non-native varieties of English. It could well be that the contact dynamics present in New English settings interact with usage frequency, yielding different outcomes.

Language contact as a crucial determinant of new varieties of English is hypothesized to influence the productivity of verb-to-noun conversion significantly. It is assumed that if a substrate language prefers verb-to-noun conversion over other nominalization processes, the blocking constraint can—to some extent at least—be overridden, which would in turn result in a higher success of V>N conversion in this variety. This scenario is envisaged for HKE and SgE, the varieties with Chinese substrata, in which V>N conversion is highly productive (cf. section 2.2).

Differences are also expected as regards the acceptability of V>N conversion. Even though V>N conversion might not be used consistently in the corpora, the Chinese substrate can be expected to still lead speakers of the respective varieties to perceive converted forms as more acceptable compared to speakers of non-Chinese substratum varieties. The judgment of the acceptability of such forms can also be influenced by the substrate. If the process is highly productive in the substrate, speakers might still be familiar with it even though it is not as productive in English. The acceptability of conversion can be hypothesized to correlate with the degree of institutionalization of English, with speakers of less institutionalized varieties accepting V>N conversion more readily.

The question in how far verb-to-noun conversion is realized differently in native and new varieties of English is explored in detail in chapter 6 and in chapter 7. Chapter 8 is dedicated to the question of acceptability.

Processing verb-to-noun conversion

In corpus-linguistic studies, it has tacitly been assumed that differences in frequencies of use of linguistic elements as represented in corpora are both the reason as well as the result of

different ways of processing these elements. It is generally hypothesized that more frequent forms are easier to process. However, this assumption is problematic considering that the non-appearance of a specific form in a corpus does not necessarily mean that it is difficult to process (cf. Schütze and Sprouse 2013: 29). For example, the adjective *carless* ('without a car') appears a mere 22 times in COCA, yet, due to its transparency, it is easily processable. It is therefore advisable to complement corpus-linguistic methods with experimental methods in order to obtain (ideally) converging results.

Whether higher corpus frequencies translate into faster speed of processing will be explored by means of a measurement of reaction times. If speakers use V>N conversion more often (as reflected in a higher frequency of occurrence in corpora) and judge it more acceptable, it is presumably also processed faster owing to the speakers' higher familiarity with the process. The more frequently speakers encounter converted forms, the more experiences they gain and the deeper entrenched the forms are in their brains. Consequently, accessing and processing these forms will be faster for these speakers compared to speakers who are not confronted with V>N conversion equally frequently. The details of how verb-to-noun conversion is processed are explored in chapter 8.

This research agenda will be addressed by adopting corpus analytic as well as experimental methods (for a detailed description of the methods cf. chapter 4), thus basing the results of the study on evidence obtained by combining two complementary research traditions. The next chapter offers an overview of previous research on the topics of conversion and word formation in World Englishes. This is complemented by a detailed critique which outlines potential stumbling blocks that are to be avoided. In chapter 3, the theoretical framework for the study is presented. Chapter 4 describes the data and methods used in this study. It presents the corpora that are analyzed and offers a critical reflection on the potential and limitations of corpus analytic studies. It further introduces the quantitative methods used, mainly collocation analysis and various types of regression modeling. Moreover, the experimental methods used in the current study are explained. Chapter 5 presents the first study. It is concerned with select case studies of the emergence of verb-to-noun conversions. Various aspects such as the development of the frequency of use as well as semantic and syntactic shifts resulting from it are the focus of this chapter. This first study only draws on data from US American English in order to lay out the foil against which verb-to-noun conversion in Asian varieties is subsequently compared. In chapter 6, corpus data from all five varieties are incorporated. It presents the results of the second study, which endeavors to compare the varieties of English from a quantitative perspective. The aim of the subsequent chapter, chapter 7, is to examine verb-to-noun conversion in Asian varieties from a qualitative perspective,

1 Introduction

analyzing a range of different aspects, among them register, constructional preferences and semantic peculiarities. The results of the corpus analysis laid out in chapters 6 and 7 are then corroborated by the results of a subsequent experiment which is presented in chapter 8. Chapter 9 seeks to connect the dots and to summarize and discuss the findings from all previous chapters.

2 Previous research

This chapter introduces concepts crucial to the analysis of conversion and gives an overview of previous research on various aspects of the topic. First, the notion of conversion is defined and contrasted with other terms that have been used to describe the same phenomenon (most notably zero-derivation). Subsequently, mechanisms and constructions potentially favoring or disfavoring conversion are described. A construction potentially favoring conversion is the LIGHT VERB construction, an antagonizing mechanism is blocking. Then it is explored in how far the substrate languages of the Asian varieties under scrutiny could facilitate or block conversion. Moreover, previous accounts of conversion in varieties of English are reviewed as well as previous attempts at addressing language contact from a usage-based perspective.

2.1 Defining conversion

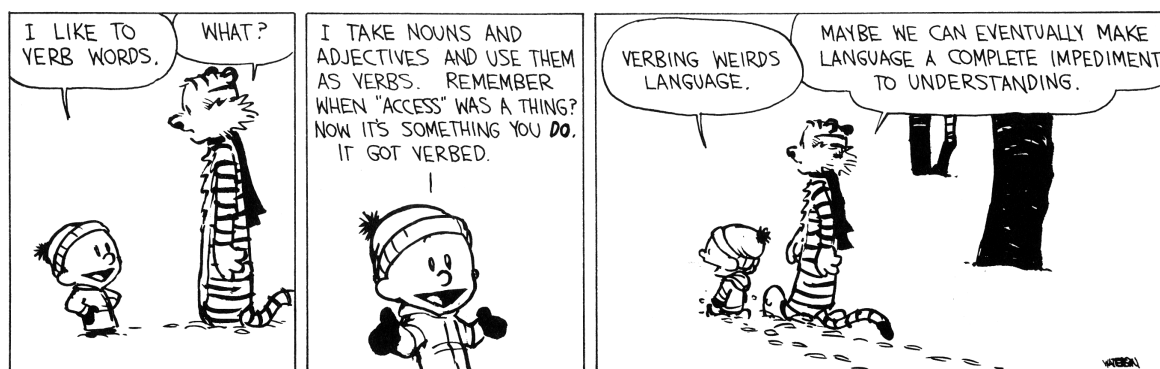


Figure 2.1: “Verbing weirds language”, Watterson (1993)

The change of word class of a lexeme without any change in its form has been labelled *conversion* (cf. Plag 2003: 107–116). When Calvin, in the above comic strip, talks about “verbing” and “weirds”, he is converting the noun *verb* and the adjective *weird* into verbs. In English, it is possible to find examples of lexemes belonging to up to five different word classes. The following sentences illustrate this for *round* (cf. Schmid 2011: 184). In 2.1 it is used as an

2 Previous research

adjective, in 2.2 as a verb, in 2.3 as a preposition, in 2.4 as an adverb, and finally as a noun in 2.5:

- (2.1) And in fact theirs has got a *round* head on it. (ICE-GB: S1B-073)
- (2.2) That's why this is to smooth them and *round* them off. (ICE-GB: S1B-043)
- (2.3) There's an interesting band *round* the walls. (ICE-GB: S2A-059)
- (2.4) Now if you come *round* here we have the Indian room... (ICE-GB: S2A-059)
- (2.5) A mass rally at Brent Magistrates Court is planned for January 10 when the next *round* of summoned offenders face the court. (ICE-GB: S2C-009)

2.1.1 Terminology

The term *conversion* was first introduced by Sweet (1891: §§105–107). Conversion is a process that has received much attention and has been looked at from many different angles. This finds expression in the quantity of terms applied to the phenomenon. The following is an overview of various approaches to conversion. Following Balteiro (2007a: 19–64), I first present those that consider the phenomenon a non-derivational process and then those that interpret conversion as a process involving derivation. Among the first group is Farrell (2001), who calls into question the notion of word class and considers so-called category underspecification a more likely scenario. That is, words are not associated with one word class only but rather have the potential to adopt any word class in a specific syntactic environment. In other words, what is underspecified is the word class (i.e. category) a lexical item belongs to. Another approach that disregards derivational processes is multifunctionality (cf. Koziol 1937: 201; Zandvoort 1972: 265). In contrast to the preceding theory of category underspecification, this approach posits that words show *a priori* multiple class membership, i.e. belong to more than one word class.

Among the approaches favoring a derivational view range those which conceptualize the phenomenon as zero-derivation or conversion. The zero-derivation approach, as proposed by Marchand (1960: 293–308) and Kastovsky (1982: 172–175), suggests that the new lexeme is formed by means of affixation of a zero-morpheme $\{\emptyset\}$, a morpheme that has no phonological weight as exemplified in 2.6:

- (2.6) *walk* (V) + \emptyset > *walk* (N)

The notion of zero-derivation is based on what Sanders (1988: 156) calls the *overt analogue criterion*. The idea behind it is to draw an analogy to inflection, where the difference between e.g. *sheep* (sg.) and *sheep* (pl.) is a purported zero element that is added to the singular form to produce a plural form. This process is said to apply to word formation as well. This means that whenever there is an evident parallel between word-formation processes, there must be some morphological material involved so as to explain the analogy. This is exemplified in 2.7 (examples taken from Schmid 2011: 189).

	legal	-ize	>	legalize	‘make legal’
(2.7)	sterile	-ize	>	sterilize	‘make steril’
	clean	-∅	>	clean	‘make clean’
	tidy	-∅	>	tidy	‘make tidy’

The criticism brought forward against zero-derivation is extensive. To further the long-standing discussion of zero-derivation versus conversion is not within the scope of this work, which is why the most striking points of criticism shall only be mentioned briefly. According to Aronoff (1976: 71),

the concept of a formless phonological substance (i.e. zero morpheme) [...] is abhorrent, even ridiculous when we realise that for every word-formation rule which has no associated phonological operation [...] we must posit [...] such entity, with a resulting *proliferation of zeroes*, one for every rule [cf. examples 2.1–2.5, emphasis added].

Bauer (2003: 38) further points out that “[e]ven if this state of affairs is possible within a generative theory of morphology, it does not have much plausibility as an account of the way in which real speakers process language”. Furthermore, the multitude of zero-allomorphs stemming from the “multiplicity of distinct analogies [...] may even suggest contrary or contradictory relations between the elements of [a verb-noun] pair” (Balteiro 2007a: 27). Finally, there is always the problematic option that “overt analogues for cases that intuitively constitute clear cases of zero-derivation” are absent (ibid.: 28), as might be the case with verbs derived from proper nouns (as in *Google* ‘name of a search engine’ + {∅} ‘??’ > *to google* ‘search for on the internet’).

This study thus rejects the notion of zero-derivation and related approaches (cf. Don 1993; Myers 1984) and adopts the term *conversion*. The term implies a derivational process (vs. category indeterminacy) that is not based on morphemes (vs. zero-derivation). Conversion is “the extension of the functional potential of a particular lexical unit beyond the limits of its word-class” (Balteiro 2007a: 34). As far as morphological structure is concerned, “no formal alternation of the original lexical unit takes place” (ibid.: 35).

The reason why here the most neutral term, conversion, is preferred, is that most of the other terms used to designate conversion (e.g. zero-derivation, category-changing inflection, functional change, functional shift, transposition) stress a particular aspect of the phenomenon, such as its syntactic potential or the parallels with morphology. Yet, a more neutral term is preferable, considering that the question with what linguistic domain conversion is associated has no straightforward answer. Conversion is a phenomenon that can be considered to belong to various domains of linguistics, as Balteiro (2007b: 15) notes: “[T]he inclusion of the so-called conversion in either Morphology [sic] or Word-Formation [sic, or syntax] depends on how those [...] disciplines are understood, but also on how the phenomenon of conversion itself is defined.” Zero-derivationalists might consider zero-derivation to be a part of word formation, analogous to affixation. Approaches that emphasize the shift in syntactic function might see functional shift as a part of syntax (cf. *ibid.*). Some approaches go so far as to consider conversion a part of inflection (cf. Myers 1984). The difficulty of assigning conversion to a linguistic domain seems to be rooted in the fact that it is non-prototypical. It is generally not regarded as a part of grammar (except for some select views such as Myers 1984), since it involves no bound grammatical morphemes; nothing is inflected. Neither is it part of traditional approaches to syntax. Consequently, conversion is often situated within word formation. Bauer (2003: 124) considers it a non-prototypical word-formation process. He stresses that there is “a central core” of word-formation processes of which conversion is not a part. According to Bauer (*ibid.*: 124–125), the most prototypical processes are of a morphological nature (“prefixation, suffixation, backformation and neo-classical compounding”). If conversion is understood as a non-prototypical word-formation process, it should receive attention from the domain of lexicology as well. Nonetheless, it remains a side issue in one of the most representative dictionaries of the English language. A cursory glance reveals that whereas the transparent construction *carless* (adj.) receives its own entry in the Oxford English Dictionary, the lexicalized verb *to holiday* does not have an entry, but only occurs in the entry for *carless* in the inflected form of *holidaying*. This aptly illustrates the difficulties of assigning conversion to a linguistic domain.

The status of conversion is thus ambiguous; it hovers between morphosyntax and word formation (cf. Plag 2003: 114). According to Plag (*ibid.*: 114–116), conversion is a lexical process because it allows for idiosyncrasies such as unclear constraints on what can be converted (e.g. *to winter* vs. **to autumn*). Furthermore, the process often leads to new elements with non-compositional, lexicalized meanings. On the other hand, some conversions are non-idiosyncratic and can be applied more rigorously. These are then rather syntactic, for syntactic processes are generally “rather exceptionless”, according to Plag (*ibid.*: 115). An

example is the adjective-to-noun conversion, by which every adjective can be converted to a noun and become the head of a noun phrase as in *the poor, the rich*. Consequently, assigning conversion to either the morphosyntactic or the word-formation domain would be too simplistic. Therefore, in this study, a Construction Grammar approach to conversion is presented, which aims to bridge the grammar-lexicon divide.

2.1.2 Productivity and constraints on conversion

Conversion is a phenomenon that is extremely productive in many languages. In their survey of *Word-Formation in the World's Languages*, Štekauer et al. (2012: 309) state that roughly 60% of the 55 languages they studied use conversion as a productive process. One of the languages which makes “unusually” extensive use of this process is English (Schmid 2011: 184). The process is so frequent that some even argue that there are no constraints on it at all (cf. Bauer 2002: 226)¹. The current state of high productivity is the result of several diachronic processes (cf. Schmid 2011: 185–186). The first is the loss of inflections that occurred between the stages of Old and Middle English. The Modern English word *love*, for example, goes back to the Old English stem *luf-* and its inflected forms *lufu* (N) and *lufian* (V). The loss of inflections has led to OE. *lufu* and OE. *lufian* collapsing into ModE. *love*. The second process is extensive borrowing of “already formally more or less identical” forms, particularly from French, in the Middle English period. The loss of inflections contributed to these lexemes collapsing into one homophonous lexeme within a short period of time from the moment of their entry into the English language. In Early Modern English, deriving words from one another became a popular means of word formation,² which is the third reason for its productivity today. Finally, over the course of the centuries, phonetic processes have led to instances of phonological merger. One example is OE. *hatian* and OE. *hete* both resulting in ModE. *hate*; a process that was probably influenced by Old Norse (cf. *ibid.*: 186).

The following sections serve to briefly introduce the notions of productivity and blocking, two key concepts in word formation.

Productivity

The notion of productivity has received much attention, yet, depending on the theoretical framework in which one's work is grounded, e.g. generativism or the usage-based paradigm, productivity is conceptualized in different ways. Consequently, there is no unified approach

¹The only restriction that he mentions is blocking (see below).

²Shakespeare made use of conversion quite frequently, e.g. “grace me no grace, nor uncle me no uncle” from *Richard II* (cf. Cannon 1985: 415).

2 Previous research

to measuring productivity. This section can only provide a brief overview of productivity as seen from a usage-based perspective; for a more detailed account the reader is referred to, for example, Baayen (2009), Bauer (2001), and Plag (1999).

Bauer (2001: 98) defines productivity as the “potential” of a word-formation process “for repetitive non-creative [...] coining”, that is, for yielding new words that have not been coined to be purposefully creative as could be expected in marketing, for example. However, productivity of a morphological process “may be subject unpredictably to extra-systemic factors”. This definition implies that productivity is inherent in the language system, a claim with which Baayen (2009: 917) disagrees on grounds of empirical evidence: “Contrary to what Bauer suggests, recent research has shown not only that the effects of ‘extra-systemic’ factors are truly predictive for productivity, but also that the ‘intra-systemic’ factors are part of a much larger system of interacting factors.” As Baayen (*ibid.*) illustrates, “historical, stylistic, onomasiological, and cognitive factors” all contribute to the productivity of a morphological process.

The main issues which remain unresolved when it comes to productivity are identified by Baayen (*cf. ibid.*: 900) as follows.

A first key question in productivity research is what conditions need to be met for a rule to be productive in these ways. A second key question is whether a rule is ever totally unproductive, i.e., whether productivity is in essence a graded phenomenon. [...] A third set of questions addresses how productivity changes through time [...]. A final issue is the relation between productivity and processing constraints in the mental lexicon.

In order to investigate these points empirically by means of corpus-based studies of language, Baayen (*ibid.*) describes three different measures of productivity which focus on different aspects:

realized productivity Also called type frequency, describes how productive a morphological category or process has been to date.

expanding productivity Is “the rate at which a morphological category is expanding and attracting new members”. Expanding productivity is measured by the number of hapax legomena belonging to this category in the entire corpus.

potential productivity Is a measure of how productive the morphological category can be expected to be. Potential productivity is calculated by dividing the number of hapax legomena of a morphological category by the total number of tokens of that category in a corpus.

The main advantage of realized productivity (and probably the reason why many studies work with type frequencies) is that it is comparatively easy to calculate. Nonetheless, realized productivity only measures past productivity and fails to reflect synchronic processing constraints. Baayen (cf. *ibid.*: 906) illustrates this by drawing on the Dutch verb prefix {VER-} compared to the suffix {-STER}, used to designate a female agent noun. While {VER-} shows a much higher type frequency, {-STER} is in actual fact the more productive affix, considering that in an experiment participants did much better at creating neologisms with {-STER}.

It can therefore be preferable to complement mere realized productivity with expanding or potential productivity. Both measures are synchronic in focus in that they are calculated on the basis of hapax legomena in a corpus. However, as Baayen (cf. *ibid.*: 904–905) himself points out, calculating the number of hapax legomena as an indicator for the number of neologisms formed by a morphological process is not without problems. While the number of neologisms can be expected to increase with corpus size, the opposite is true for the number of hapax legomena: with an increase in corpus size the number of hapax legomena is expected to decrease. The relation between the number of hapax legomena or the number of types and the total size of the corpus is thus not linear. Säily and Suomela (2009), for example, describe a method to compare the productivity of a morphological process across corpora of different sizes which takes this nonlinearity into account.

This study relies on token frequencies of a fixed number of types, a measure which comes closest to Baayen's realized productivity. While making claims about productivity on the basis of only token frequencies is highly error-prone (see the conserving effect of highly frequent forms such as irregular verbs which are usually formed by means of fairly unproductive processes), it makes sense to use this measure in the present corpus study, as the number of types is fixed (cf. chapter 6). As pointed out above, this measure can only capture past productivity without predicting any future development (as expanding or potential productivity would do). Yet, it is the most robust measure in the present case. Relying on hapax legomena, as would be required when measuring the productivity of conversion by means of expanding or potential productivity, is not recommendable with the data at hand. As will become clear in chapter 4, the corpus used for this study provides a very large but also very 'messy' extract of the web. As such the language is in part of a conceptually spoken nature, which leads to the data being highly susceptible to mistakes, which in turn makes a reliable identification and classification of hapax legomena very difficult.

The blocking constraint as an effect of frequency

Despite the fact that conversion is so frequent and pervasive in Modern English, there do seem to be some examples of conversion that are somehow more remarkable than others. These remarkable instances of conversion are the object of analysis in the present study. Their noteworthiness lies in three key aspects. The first is directionality. While there are numerous accounts of noun-to-verb conversion for the English language (cf. e.g. Clark and Clark 1979; Dirven 1999; Karius 1985; Marchand 1969; Zandvoort 1972; for an overview cf. Baeskow 2006), specific accounts of verb-to-noun conversion are rare, which is most likely due to the much higher productivity of verbalizations (e.g. *a key* > *to key a message*, cf. Don et al. 2000: 949).³ Thus, any conversion that is off the beaten path of noun > verb will be of higher salience.

The second aspect is the degree of conversion. There is a tradition to distinguish between full and partial conversion (cf. Sweet 1891: §§106–107).⁴ Full conversion has occurred when the resulting lexeme has “adopt[ed] all the formal characteristics (inflection, etc.) of the part of speech it has been made into” (ibid.: 39), e.g. when a noun that results from verb-to-noun conversion is no longer confined to restricted environments with “semantically empty verbs” such as [*have a N*] or [*take a N*] (Balteiro 2007a: 50), but shows a plural morpheme and is freely modifiable by adjectives etc. Converted forms in subject position can be assumed to have reached a very high degree of ‘nouniness’, considering that the noun phrase and the subject function are prototypically nominal. However, full conversion where converted forms occur in subject position is comparatively rare, as is evident from, for example, Marchand (1960: 304), who notices that only 11% of his data points “show the deverbal sbs [= nouns] as subject of the sentence”.

The third aspect regards constraints on conversion. Bauer (2001: 126) points out that the term *constraint* indicates “that the restrictions are not necessarily absolute” and can therefore be violated. Whenever a constraint is violated, the instance of conversion is likely to attract more attention than constraint-conforming formations. There is no one clear account of constraints on word formation,⁵ but the common ground in all works on constraints is the existence of a blocking constraint, which also applies to the process of conversion. Blocking

³Schmid (2011: 199) attributes the low frequency of verb-to-noun conversions to the fact that “deverbal nouns are based on the relationship ‘whole for part’, e.g. ACTION for OUTCOME”. He goes on to state that “[m]etonymies based on this relationship [...] appear to be conceptually less helpful and productive” and are therefore comparatively infrequent.

⁴See Balteiro (2001: 10–11) for a critical account of partial conversion. She goes so far as to reject the notion of partial conversion entirely.

⁵For an extensive discussion of constraints that have been proposed for morphological processes, cf. for example Bauer (2001: 126–143) and Plag (1999: 37–61). Most of these constraints refer to formal characteristics of

implies that frequencies of occurrence of rival patterns, e.g. conversion and derivation, can influence each other. Aronoff (1976: 43) describes it as “the nonoccurrence of one form due to the simple existence of another”, so as to avoid the creation of synonyms. An example of this constraint is the non-existence of the word **stealer* due to the existence of *thief* in English (cf. e.g. Bauer 2001: 136). In Construction Grammar, this principle is called *statistical preemption* (Boyd and Goldberg 2011).⁶ Hilpert (2014a: 138–139) explains it as follows:

First, speakers form generalisations over sets of constructions that are comparable with regard to their meanings. [...] Second, it is assumed that speakers keep a detailed record of the lexical elements that they hear in these constructions. [...] Consequently, when] speakers perceive a statistical imbalance, [...] they interpret that imbalance as meaningful: if a lexical item rarely or never appears where it would be expected with a certain base frequency, then it is absent because of a constructional constraint.

Boyd and Goldberg (2011: 80) were able to show that in an experimental setting, subjects could be induced to infer constructional constraints. They attribute this to the subjects’ reliance on meaningful language production from the speaker (cf. Cooperative Principle, Grice 1975). This means that “hearer[s] construct[.] an explanation” by way of their ability to “mind-read” (Hilpert 2014a: 141). The conclusion that Boyd and Goldberg (2011: 80) draw from this is that “categorization and statistical preemption play a role in restricting linguistic productivity”. In other words, skewed input frequencies lead to a preference of one construction over another, semantically similar construction.

Particularly in word formation, statistical preemption is mainly referred to as the blocking constraint (cf. Aronoff 1976: 43). Of two word-formation processes yielding potentially synonymous forms, the one that is more frequent for a specific lexical item will most likely be the preferred one, thus blocking a rival mechanism that would yield a (near-)synonym. Hilpert (2014b), for instance, shows this preference for one construction over the other for adjectives ending in *-ic* and *-ical*. For deverbal nouns, two such processes that can potentially block each other are conversion and derivation. To give an example, for the deverbal noun to describe the process of ‘inviting’ suffixation (*invitation*) is usually the preferred option. In COCA, the lemma *invitation*⁷ yields 7465 hits. The semantically similar but converted form, *invite*⁸, only yields 217 tokens. For the lemma *increase*, on the other hand, conversion is

derivational affixes and the bases they are combined with (e.g. affix ordering) and are therefore of no further interest at this point.

⁶In the present context, *blocking* and *statistical preemption* are used to refer to the same process, as blocking is understood as a special case of statistical preemption.

⁷[invitation].[n*]

⁸[invite].[n*]

the preferred option. COCA yields 35,781 tokens for the converted form⁹; the synonymous derivation *increasement*¹⁰ is marginalized to a mere two tokens.

Nevertheless, statistical preemption can change over time when input frequencies are modified. The “statistical imbalance” (Hilpert 2014a: 139) that may have led to constructions blocking one another can even out and consequently give way to new constraints. If these new constraints favor a construction that has not occurred very frequently up to that point, this construction might gain momentum and establish itself alongside the other construction. Plag (1999: 52) explains statistical preemption in word formation as follows: “[I]n order to be able to block a potential synonymous formation, a word must be sufficiently frequent”. However, if it does not occur with sufficient frequency, the potential synonym can rise in frequency and spread. Rainer (1988: 164) elaborates on the blocking constraint:

[W]e may view the blocking force as the result of the antagonism between the pressure exerted by a potential regular word and the resistance offered by the corresponding blocking word, whereby pressure is a function of productivity and resistance a function of frequency.

Productivity is ultimately frequency, moderated by statistical experiences that speakers may have with a particular construction. In the subsequent study it will be shown that statistical preemption is crucial in shaping the usage patterns for verb-to-noun conversion in US English (cf. chapter 5). Furthermore, chapter 6 will illustrate that the blocking constraint may apply to different degrees in different varieties of English, inducing distinctive usage patterns of conversion.

2.1.3 Light-verb constructions

In the context of conversion, the LIGHT-VERB construction (LVC) is worth a closer look. LVCs are defined as consisting of a semantically bleached verb and another verb that carries the lexical content and has been converted to a noun, as in 2.8 (cf. Dixon 2005: 459–483).

(2.8) Mary **had a walk** in the garden.

Verbs that qualify as semantically bleached verbs are *have*, *take*, and *give*, and, to a lesser extent, also *make*, *do*, and *pay* (cf. *ibid.*: 459, 461). The full verb is preceded by an indefinite article. Whether the converted form can be preceded by a premodifier as in *Mary had a long walk in the garden* is debated, with e.g. Dixon (*ibid.*: 464–465) arguing for and Hoffmann

⁹[increase].[n*]

¹⁰[increasement].[n*]

et al. (2011: 267) and Wierzbicka (1982: 755–756) against it. The main argument against premodification is that the semantic equivalence between the simplex form and the light-verb construction is no longer intact once a premodifier is inserted (cf. Hoffmann et al. 2011: 267). The same argument holds for postmodification. The LVC thus has the following form: $[V_{\text{light}} a V_{\text{lexical}}]$.

The occurrence of particular verbs in LVCs is generally semantically motivated (cf. Dixon 2005: 460) and each of the light-verb frames shows its own semantics.¹¹ A general characteristic of the construction is its association with the colloquial register, which is evident in the impossibility of formulations such as **to have a urinate* compared to *to have a pee* (cf. *ibid.*: 461, 483).

LVCs can potentially facilitate conversion because the converted form does not have to adopt all the characteristics typical of nouns. As Dixon (*ibid.*: 466) claims, the lexical verbs used in LVCs are not to be confounded with established conversions.

LVCs seem to be of particular relevance in the study of conversion in new varieties of English. Hoffmann et al. (2011), in a study of the SAVE corpus, a corpus of newspaper articles from South Asia, find that LVCs are most common in IndE compared to other South Asian varieties. They attribute this to the high degree of institutionalization of IndE. While LVCs are associated with the colloquial register in the parent variety (British English), they have made it into the more formal newspaper register in IndE due to the growing endonormativity that comes with the increased indigenization of English in India. Bernaisch (2015: 170–193) also finds that IndE shows markedly distinct usage patterns for LVCs. This development could probably also be observed for other indigenized/indigenizing varieties of English. In the qualitative analysis of conversion in Asian Englishes (cf. section 7.4.3) the productivity of LVCs is analyzed in detail.

2.2 Conversion in the substrates

In order to understand in how far substrata could potentially play a crucial part in shaping the usage pattern of conversion in Asian varieties of English, it is adamant to examine them more closely. Throughout this study, the term *substrate/substratum* is adopted to refer to the languages that have come in contact with English. While these languages are technically adstrates to the new varieties of English, they were substrates to English when the British

¹¹For detailed accounts of the semantics of LVCs, the reader is referred to Wierzbicka (1982) and Dixon (2005: 469–476).

settlers first arrived. From a historical perspective it is therefore appropriate to talk of substrates.

2.2.1 Conversion in Chinese dialects

A variety of Chinese dialects¹² have come in contact with English to yield Hong Kong English and Singapore English; the two major dialects being Mandarin and Cantonese. The following is based on Matthews and Yip's (1994: 5) premise that Mandarin and Cantonese "are not mutually intelligible" but that "their grammatical structure is similar in most major respects". In the case of verb-to-noun conversion, Mandarin and Cantonese are structurally the same (Bao Zhiming p. c., July 9, 2014). Therefore, no distinction between Mandarin and Cantonese is drawn.

As regards verbs and nouns, Chinese is truly a language of category indeterminacy. Using verbs in nominal context is a comparatively unconstrained process in Chinese. This process is facilitated by the fact that in Chinese, a typologically analytic language, words are not inflected (cf. Ross and Ma 2006: 22). Thus, according to Po-Ching and Rimmington (2004: 16),

[n]ominalisation in Chinese does not usually seek morphological conversions. It is always context-dependent. In other words, all nominalisations are contextual nominalisations. A verb or an adjective may be taken as a noun therefore [...] in a given context or grammatical framework.

These "given context[s]" seem very broad compared to English, as this statement by Matthews and Yip (1994: 55) shows: "[W]hile any verb in Cantonese can appear in subject and object positions without change in form, verbs in English generally take affixes if they are to appear in these positions". The following examples (Bao Zhiming p. c., July 9, 2014) illustrate the verbal and nominal use of the verb *choose* in Chinese.

(2.9) *wo xuan le zhenque de ke*
I choose ASP correct PART class
'I chose the right class.'

(2.10) *wo zuo le zhenque de xuanze*
I make ASP correct PART choice
'I made the right choice.'

¹²See Leimgruber (2013c: 3) for an explanation as to why varieties of Chinese are usually called dialects. I am herein adhering to this nomenclature.

In both examples, the past tense is expressed with the help of the aspectual particle *le*. The particle *de* is an indicator that the preceding element (*zhenque*) is a premodifier to the following element (*ke* and *xuanze*, respectively). As is easily visible, the form of *xuan* does not change when it is used in a nominal context such as example 2.10. Yet, *xuan* is modified to *xuanze*, which is, however, a compound rather than a noun and a suffixed morphological marker. In Chinese, in order to facilitate the task of distinguishing between the verbal and the nominal use of the word, verbs are generally monosyllabic whereas nouns tend to be bisyllabic. To fit this pattern, *ze* is inserted after *xuan* in 2.10, the nominal context. *Ze* is a synonym of *xuan*, not a morphological marker as used in English derivational processes (e.g. {-NESS}, {-TION}). What Chinese does to comply with bisyllabicity in nouns is to reduplicate the meaning, which means that pattern fit comes at the expense of an increase in redundancy (Bao Zhiming p. c., July 9, 2014). The semantically explicit interpretation of example 2.10 would consequently be as follows, where *xuan* and *ze* mean the same:

- (2.11) *wo zuo le zhenque de xuanze*
 I make ASP correct PART choice-choice
 ‘I made the right choice.’

This example shows that, except for the tendency to prefer bisyllabic nouns, Chinese words can easily be used in verbal and nominal contexts without a change in form. This interchangeability is restricted for other cases, as Po-Ching and Rimmington (2004: 16) explicitly points out: “[o]ther word classes [i.e. other than verbs and adjectives] are less likely to become nominalised”. For example, the highly productive noun-to-verb conversion in English (e.g. *access* > *to access*) is infrequent in Modern Chinese (cf. Shi 2006: 309; Bao Zhiming p. c., July 9, 2014). In Chinese, the wall is not ‘painted’, but ‘paint is applied to the wall’. This could be due to a general preference for explicitness in Chinese verbs. A bag is not carried, but carried on the shoulder (*bei*), with the hand (*ti*), in one’s arms (*bao*) or under the arm (*jia*; Wei Chen p. c., July 6, 2014). This could explain why noun-to-verb conversion is of very low productivity in Modern Chinese. Only extended and intensive language contact between English and Chinese, as is the case in Singapore, has led to this widespread process seeping into Chinese from English, so that Singaporeans, when speaking Chinese, might eventually ‘color a book’ due to their increased exposure to English and English conversion processes (Bao Zhiming p. c., July 9, 2014). Shi (ibid.) describes the same contact phenomenon in Hong Kong written Chinese.

Verb-to-noun conversion is thus an ideal starting point for the analysis of Chinese substrate influence in contact varieties of English. Since the Chinese substrata allow for verbs to be used in nominal position in virtually every context and without any change in form, one

is likely to find traces of the ease with which verbs convert to nouns in varieties of English with a Chinese substratum. The fact that Chinese nouns tend to be bisyllabic and speakers, in forming nouns, draw on reduplication is not expected to impede conversion. In English, reduplication is a process of extremely low productivity. Consequently, it is unlikely that speakers transfer the reduplication pattern.¹³

As has been pointed out, verb-to-noun conversion in English is moderately frequent, examples such as *an invite* or *the disconnect* are common. Nonetheless, some formations seem to be blocked (e.g. **the receive*). In contact with Chinese, however, the productivity of V>N conversion might experience an increase in type frequency and thus also an increase in token frequency, leading to the spread of such purportedly illicit formations.

2.2.2 Conversion in Malay

According to the 2010 census, 12.2% of the Singaporean population speak Malay at home (cf. Wong 2011). In order to adequately compare SgE to HKE on the grounds of the structural properties of the substrate languages, it is necessary to briefly touch on the process of conversion in Malay, the second-largest substrate in Singapore after Chinese.

Malay is an agglutinating language and as such shows no inflection (cf. Maxwell 1907: 45). Furthermore, many words can appear in various different word classes depending on the context (cf. e.g. Crawford 1852: 43; Knowles and Don 2003: 422):

A difficulty which attends the classification of Malay words into various parts of speech, according to the system applied to European languages, consists in the number of words which, while yet unmodified by particles, are either verb or substantive, substantive or adjective, adjective or adverb, *according to the context*. [...] The same thing occurs in English in a minor degree; [...] Many Malay words must thus be treated as now substantive, now adjective, now verb, according to the position they occupy in the sentence. (Maxwell 1907: 45, emphasis added)

Thus, the situation in Malay is similar to that in Chinese, where the context is an important determinant of the word class with which a particular word is associated. Owing to the similarity of Malay and Chinese in this respect, it can be assumed that a comparison of the process of conversion in SgE and HKE on the grounds of similar characteristics of their substrate language/s is valid.

¹³In line with Bao (2005, 2009, 2010a, 2010b), it is hypothesized that the lexifier language serves as a filter and that it will “flush out” the reduplication pattern during the transfer process due to its incompatibility with the English superstratum (cf. section 2.4).

2.2.3 Conversion in Hindi

To analyze the magnitude of the influence of the Chinese substratum on English, another Asian variety with a typologically different substratum has been chosen as a basis of comparison. If an analytic substratum in which verb-to-noun conversion is unconstrained fosters the process in the contact variety, this is not expected to happen to the same extent in a contact variety of English with a synthetic substratum. The Asian variety selected for comparison is Indian English. It has a multitude of adstrata as figure 1.7 on page 22 shows. However, the main contact language of English in India is Hindi, which is a synthetic language.¹⁴

In Hindi, suffixation is the preferred process for nominalization (cf. Kachru 2006: 111–118). Hindi has a plethora of derivational suffixes that can attach to verbal roots. One of them, the infinitival {-NA}, combined with a verbal root, yields a form that can function as either the infinitive or an abstract noun. An example is *gana*, which can either mean ‘to sing’ or ‘song’ (cf. *ibid.*: 115). However, Kachru (*ibid.*: 114–116) lists nine other ways of deriving abstract nouns from verbs by means of suffixation so that it can be assumed that the suffixation by way of adding {-NA} to the verbal root and thus yielding a form that is identical to the infinitive is not as unconstrained as the possibility of using any verb in a nominal slot in Chinese. All other types of nouns in Hindi, e.g. “concrete nouns” or “action nouns” are derived by means of other suffixes (e.g. *k^helna* ‘to play’ > *k^hilɔna* ‘toy’, *k^hana-pīna* ‘to eat-drink’ > *k^han-pan* ‘food and drink’, *ibid.*: 118).

2.3 Conversion in World Englishes

The topic of word formation in World Englishes in general and conversion in particular has so far not received much attention. An early study on conversion in varieties of English comes from Cannon (1985), focussing on US American English. The database he draws on is a collection of dictionaries. The two main methodological challenges he identifies in his own study are, firstly, that he investigates written material only and consequently fails to capture conversion in spoken language. Secondly, that the written material he uses are dictionaries, that is, heavily edited documents of language. Furthermore, dictionaries do not provide frequency data, but explicitly make a point in omitting forms that are too infrequent

¹⁴Considering that all languages other than Hindi are spoken by under 10% of the population (cf. figure 1.7), a thorough investigation of their properties as regards conversion is deemed unnecessary.

(cf. *ibid.*: 413). Nonetheless, Cannon (1985: 416) finds a total of 567 “functional shifts” and makes various claims about frequent directions and registers in which conversion occurs.¹⁵

The most exhaustive study on word formation in New Englishes so far has been conducted by Biermeier (2008). The data for it were extracted from the *International Corpus of English* (ICE). Of the varieties explored in this project, Biermeier includes British, Indian, and Singapore English. For these varieties, Biermeier (*ibid.*: 97) finds 743, 571 and 754 tokens respectively, which, according to Biermeier, indicates that BrE and SgE show similar levels of productivity for conversion and that in IndE conversion is less productive. As far as the direction from verb to noun is considered, the investigation yields very similar type frequencies across the three varieties (cf. *ibid.*: 87). As regards registers, his results are that in SgE, conversion occurs in a “distinctly vast range of texts”. In IndE, on the other hand, conversion is more restricted to written texts (cf. *ibid.*: 99, 90). He explains the high frequency of conversion in SgE with the high proficiency of the speakers; they are “very close to the native status” (*ibid.*: 88). For IndE, the fact that the comparatively low token frequency contrasts with four new types that occur in none of the other varieties (probably indicative of a higher productivity of the process) remains largely unexplained (cf. *ibid.*: 99).

After this brief exposition of results, it is necessary to cast an eye on some methodological challenges that Biermeier’s (*ibid.*) study faces, so as to avoid them here. The first is the database. Since the ICE subcorpora only contain one million words each, Biermeier (*ibid.*: 15) struggles with quantitative analyses of the data. Haselow (2010: 134–135) in a review of Biermeier’s (2008) research even goes so far as to suggest that “many conclusions made in the book are rather doubtful” and that “the analysis is often restricted to a mere verbalization of frequency phenomena rather than offering abstractions and an indication of general tendencies”. Nonetheless, Biermeier’s study has its merits in that it illustrates the difficulty of pinpointing divergent tendencies in World Englishes. It is highly unlikely that varieties will drop one word-formation process entirely or develop new, unprecedented usage patterns. Rather, it is to be expected that differences between varieties are gradient in nature (cf. Schneider 2007: 80). These subtle differences are, however, difficult to grasp with the comparatively small amount of data that Biermeier (2008) draws on (also cf. section 4.1 on the limitations of ICE due to its size). The publication of the *Corpus of Global Web-based English* (Davies 2013) in 2013 can therefore be seen as a lucky chance for research on word formation in World Englishes. Larger amounts of data will yield a higher number of relevant tokens and will thus facilitate the task of drawing overarching conclusions.

¹⁵Since this research focuses mostly on Asian varieties of English, a detailed account of Cannon’s results will not be given here.

Another methodological challenge with which Biermeier's investigation of conversion is confronted is that in order to evaluate conversion it relies on a fixed type list. On this list are well-established conversion pairs such as *arrest*, *attempt*, *smile*, *whisper* and the like. While it is possible to establish the direction of conversion in these cases with the help of etymological dictionaries, these are instances of long-standing full conversions. From a cognitive perspective, these formations cannot be considered to be of the same type as novel conversions such as */*an imagine* or */*a develop*. Whereas in the latter cases, native speakers could easily establish the direction of conversion and would most likely judge the novel formations as ungrammatical, the same is incomparably more difficult in cases of established conversions. Balteiro (2007a: 39) calls instances of conversion where the direction of conversion cannot be established and the forms in question have become well-entrenched as members of at least two word classes "pseudo-conversion[s]". She notices that distinguishing true from pseudo-conversion "is only important for the linguist. For language users, it is, however, completely irrelevant." She goes on to say that all "synchronically identical word pair[s] which [are] semantically related" can be regarded as instances of multiple word class membership, regardless of their historic trajectory. For the language user, establishing directionality is impossible. Furthermore, because in cases such as *love* or *smile* the verbal and nominal form are both very well entrenched, none of the forms is likely to be judged as ungrammatical by language users. In order to establish in how far a contact language can influence the English word-formation process of conversion, innovations, i.e. formations that have not been attested in standard varieties of English, will have to be in focus. It is particularly these formations that a speaker of an ESL variety of English might find acceptable, whereas a speaker of a native variety might judge them as ungrammatical. Haselow (2010: 133) concludes that Biermeier's data "do not allow for conclusions on the lexical creativity of the speakers or writers of a given variety, but rather document the frequency of occurrence of word-forms that *have already become established* in standard varieties [emphasis added]". The approach that will consequently be pursued in the present study is more inductive in nature in that it focuses exclusively on forms that have not been attested in dictionaries of standard varieties of English.

That the reliance on pre-fabricated word lists is not an ideal starting point is also evident in Evans's (2015b) study on word formation in Hong Kong English. He notes that around 20% of the items on the vocabulary list on which he bases his work do not occur in the corpora he investigates (cf. *ibid.*: 125). Considering that one of the corpora he analyzes is the HK section of the *Corpus of Global Web-based English*, which contains over 40 million tokens (cf. section 4.1.3), it can hardly be argued that the words in question form part of the HKE lexicon.

A further shortcoming of Biermeier's (2008) study is that it does not provide a qualitative analysis of the data. Particularly in the case of conversion, the linguistic context is of great importance to assess the status that conversion has in each variety. A study on conversion must account for instances of such a diverse nature as the following:

(2.12) Compare these skills with those students learn in today's schools and **the disconnect** is clear. (COCA, ACAD)

(2.13) Grid-connected photovoltaic systems routinely **have a disconnect** that activates when the rest of the grid goes down. (COCA, MAG)

While 2.12 is a clear example of a full conversion, 2.13 instantiates the light verb construction, a frame that could facilitate conversion. In order to account for such qualitative differences, any quantitative study of corpus data has to be accompanied by a qualitative analysis that includes the discourse-pragmatic and syntactic context of the converted forms.

2.4 Modeling frequency effects in language contact

There are numerous accounts of the development of features of new varieties of English, which are often based on explanations involving language contact and substratal influence resulting from this contact. While the work on contact-induced language change does have its merits and has helped to better understand how varieties emerge and develop, it also has to be noted that

substratist argumentation, regardless of one's theoretical persuasion, is almost solely driven by individual grammatical features that are attested in the contact language¹⁶ and can be traced to the linguistic substratum. This line of argumentation has been criticized as unsystematic and unprincipled (see Dillard 1970, Bickerton 1981, Mufwene 1990, Siegel 1999, Bao 2005). Furthermore, there has been little or no attempt to examine the productivity, as measured by frequency of use, of transferred substratum features in the contact language. (Bao 2010a: 793)

Bao (*ibid.*, also 2005, 2009, 2010b) therefore proposes a more systematic and "usage-based approach to substratum transfer" that accounts for distinct degrees of productivity and frequencies of use in language contact. This approach is of particular relevance for a study on conversion. As has already been mentioned, word formation is not a linguistic domain

¹⁶The notion of *contact language* is used by Bao to refer to the emergent variety of English, not to the language that comes into contact with English. The latter is generally referred to as the substrate language in his work.

where categorical differences between varieties are likely to develop. In a cross-varietal account of conversion it is hence necessary to be on the lookout for different usage patterns by assessing the frequencies of use of this specific construction.¹⁷

Bao's (2009: 348) model is based on two antagonistic mechanisms that are supposed to be operating at the same time. He assumes that these *constraints*, as they are called, can be violated (based on the concept of *constraints* in Optimality Theory, cf. Prince and Smolensky 2004). The constraints, named SYSTEM TRANSFER and LEXIFIER FILTER, are specified as follows (Bao 2010a: 812).

1. SYSTEM TRANSFER: Substratum transfer involves an entire grammatical subsystem.
2. LEXIFIER FILTER: Morphosyntactic exponence of the transferred system conforms to the (surface) structural requirements of the lexical-source language.

SYSTEM TRANSFER states that only entire subsystems of the grammatical system of the substratum are targeted by transfer processes. For SgE, for example, Bao (2005: 250) observes that all means of expressing perfective aspect must have been transferred from the Mandarin substratum. All variants are found in SgE, and native speakers of SgE judge them as “acceptable” (Bao 2010a: 800). The fact that not all of these possibilities surface in the new variety to the same extent—two variants show a very low frequency of use (to the point of non-occurrence), the other one is used more productively (cf. *ibid.*)—is the result of the second constraint, LEXIFIER FILTER. Only those constructions of the substratum that are compatible with the morphosyntax of the lexifier language ‘survive’ in the emerging contact language. As Bao (2009: 347) puts it:

The exponence of the transferred system is subject to the grammatical requirement of the language that provides the morphosyntactic materials, flushing out, at varying degrees of thoroughness, those elements of the transferred system that cannot be expressed felicitously.

Generally, the constraints imposed on the emergent variety by LEXIFIER FILTER rank higher than those of SYSTEM TRANSFER (cf. Bao 2005: 258), which explains why there are grammatical subsystems that only partially resurface in the new variety (such as the perfective markers in SgE).

¹⁷Bao (2010a: 794) predominantly uses the term *feature* but understands features to be constructions, i.e. form-meaning pairings in the Construction Grammar sense. For the sake of consistency, I will use the term *construction* in the following.

In shaping the emergent variety, the usage patterns of the respective constructions in the substratum and superstratum languages are decisive. It is those usage patterns that influence the productivity and thus the frequency of use of the construction in question in the contact variety (cf. Bao 2010a: 817). In this process, particularly the lexifier is of importance since it is “forms which are frequently-used [sic] in the superstratum language” upon which transferred constructions are “modeled” (Bao 2009: 349). Another factor which determines the productivity of transferred constructions is the degree to which the construction violates constraints of the emergent variety (cf. *ibid.*: 350). Bao (*ibid.*) assumes a “correlation between constraint violation and the level of productivity of the feature in the contact language”. This correlation is displayed in table 2.1 (*ibid.*: 346, 350).

Table 2.1: Correlation between violation of constraints and productivity

structures of sub- and superstratum	violation of constraints	productivity
convergent	none	normal
divergent	weak	low
divergent	strong	not productive

Strong violators are those constructions which violate “grammatical requirements” of the emerging variety. The grammatical rules of the emerging variety “may be derived from the competing languages in its contact ecology, especially the lexifier language, or emerge independently in the contact language as a result of internal drift” (Bao 2010a: 796). Constructions which violate these grammatical rules are not productive in the emergent variety.¹⁸ Constructions which do exist in the emergent variety but still violate constraints of the lexifier language are called weak violators. They show reduced productivity. This is the case for “basilectal features, if derived from the substratum” (*ibid.*). Those constructions that do not violate any constraints, neither in the emergent variety nor in the lexifier language, become most productive. The perfective aspect markers in SgE can serve as an example. All three ways of expressing perfective aspect in SgE show a corresponding construction in the substratum language, Chinese (cf. Bao 2005: 252). However, only one of the three constructions exists in English. This construction is the one that is most productive in SgE (cf. Bao 2010a: 800). The other two ways of expressing perfective aspect cannot be realized (as successfully, cf. *ibid.*) with English morphosyntactic material, and are thus “filtered out of Singapore English” (*ibid.*: 814).

¹⁸One example of strong violators are the Chinese aspect marker categories of stative imperfective and tentative, involving constructions such as *V-ing V* (*V-zhe V* in Chinese) and *V-V*. They have no equivalent in the English lexifier and are thus “filtered out” (Bao 2010a: 798).

Consequently, frequency in the contact variety can be attributed to the structural convergence of the contributing languages. Structural convergence of substratum and lexifier language leads to an increased productivity of the transferred construction. Increased productivity, in turn, leads to a higher frequency of use of the transferred construction, and a “[h]igh frequency of use facilitates the stabilization of the structure of a substratum feature” (ibid.: 817). In contrast, infrequent or non-existent substratum constructions can be traced back to structural divergence between substratum and lexifier. If there is no adequate way of expressing a construction transferred from the substratum by means of morphological and lexical material provided by the lexifier language, this construction will not surface in the emergent variety. In short, the substratum contributes the constructions and the lexifier language contributes the usage (i.e. frequency) patterns to the emergent contact variety.

According to Bao (2010b, 2011), there is a further way for the substrate to influence the emerging variety: convergence-to-substratum. This mechanism applies in the case of non-violators, that is, when English constructions have a structural equivalent in the substratum language. Convergence-to-substratum designates the process whereby “English grammatical features converg[e] in usage or function to the equivalent features in the linguistic substratum” (Bao 2010b: 1729). An example is the English modal *must*, which can have a deontic or epistemic meaning in English (cf. ibid.). In Chinese, however, “[t]he epistemic meaning is expressed lexically” and only the deontic meaning is expressed through a modal that compares to the English *must* (ibid.: 1736). Therefore, in SgE, *must* is more frequently used with a deontic meaning, contrary to what is attested for other varieties of English (cf. ibid.: 1731). Bao (ibid.: 1736) explains this comparative overuse of the deontic meaning with the convergence-to-substratum mechanism: The English construction has “acquire[d] the usage pattern[.]” of the corresponding Chinese construction. This process, contrary to substratum transfer, is “gradual” and its effects are “subtle” (ibid.).

In short, there is a general preference in usage for the construction that is most compatible with both the substratum and the lexifier. This holds for non-violating constructions such as the modal *must*, where the construction with the meaning that comes closest to the corresponding Chinese construction is preferred, as well as for the emergence of constructions in the contact variety more generally, which hinges crucially on whether the construction transferred from the substratum can be expressed by means of the lexifier. Thus, the more convergent the substratum and the lexifier language are with respect to a construction, the more productive this construction is going to be in the newly emerging variety.

Bao’s approach to language contact can also be applied to word formation, more precisely to the competition that can be expected to exist between conversion (no bound morphemes,

analogical) and derivation (bound morphemes, synthetic). Of the two rival formations the one that is structurally closer to the substratum is expected to show a higher frequency of use. Both variants do not violate the LEXIFIER FILTER constraint as they are English formation types. Nevertheless, whenever the synthetic formation type does not exist or is dispreferred in the substratum, which is the case in varieties with a Chinese substratum, this is hypothesized to lead to tensions because of divergent structures in substratum and superstratum language. The non-morphemic formation type is not expected to trigger such tensions as it exists in both the Chinese substratum as well as the English superstratum. Thus, in HKE and in SgE, conversion, the non-morphemic word-formation type, is expected to be encountered more frequently than in other varieties with largely synthetic substrate languages such as IndE.

2.5 Summary

This chapter has provided an overview of the phenomenon of conversion and of previous studies concerned with it. It has revealed that many aspects of conversion remain unclear, particularly constraints that operate on conversion as well as its productivity in different varieties of English. This study seeks to shed more light on the phenomenon of conversion in general and the aforementioned points in particular. The following chapter will lay out in how far the theoretical concepts presented in this chapter will be implemented in this study.

3 Theoretical framework

3.1 Implementing usage-based modeling in word formation

3.1.1 Conversion versus derivation

In the present study, verb-to-noun conversion is contrasted with derivation of nouns from verbs, mostly by means of suffixation. Comparing the DEVERBAL CONVERTED NOUN construction to a competing construction is quintessential when taking a usage-based approach to language, as Schmid (2015: 21) asserts:

For frequency counts of individual linguistic items to be meaningful in terms of conventionality and entrenchment, they have to be measured and interpreted *relative* to frequencies of syntagmatic companions [...], to frequencies of paradigmatic competitors, and to frequencies of pragmatic competitors [...]. [emphasis added]

Derivation can be understood as a “paradigmatic competitor” of conversion. Table 3.1 contrasts the features of the conversion process with those of the derivation process in a simplified way (features based on Williams 1987: 169–191).

Table 3.1: Conversion vs. derivation

	conversion [V] _N	derivation [V + suffix] _N
substantive cxn	?	+
schematic cxn	?	+
atomic cxn	+	–
complex cxn	–	+
regularization	+	–
redundancy	–	+
ambiguity	+	–
processing effort	+	–

3 Theoretical framework

Conversion results in seemingly atomic, substantive constructions but is itself a schematic process, whereas derivation is both schematic and also substantive and yields complex constructions. Due to the absence of morphological material, conversion can be said to be regularizing, to reduce redundancy. At the same time, because the verb and the deverbal noun have the same form, conversion increases ambiguity which leads to higher processing efforts (e.g. those efforts that coercion requires). In order to reduce processing costs in conversion, conversion can be embedded in unambiguously nominal contexts. This is likely to happen considering that

[t]he more easily speakers can identify the parts of a construction [e.g. the NOUN PHRASE construction], the more easily that construction will accommodate other constructions [e.g. the DEVERBAL NOUN construction] into its open slots (Hilpert 2014a: 93).

Thus, in those cases where ambiguity and, with it, the processing effort are to be reduced to a minimum, conversion is likely to occur within strongly entrenched constructions like the NOUN PHRASE construction.¹

3.1.2 Chunks and lexical diffusion

Chunking is an essential mechanism in language processing and production. Contrary to what could be assumed, in natural language processing, language is usually neither produced nor processed word by word (even though speakers can be induced to process language word by word, for example in an experimental setting such as the maze task presented in chapter 8). It is rather the case that language is perceived in building blocks, so-called chunks. Chunks, however, are not random clusters of words. Words which co-occur frequently are constituent parts of chunks. Schneider (2014c: 2) defines chunks as “mentally represented multi-word unit[s]”. Their size is not limited to a certain number of words.

There is considerable evidence for the existence of chunks. For language perception, Arnon and Snider (2010: 76) for example find that four-word expressions occurring more frequently are perceived faster than those occurring less frequently. For production, Janssen and Barber (2012: 10) find that more frequent two-word expressions are named faster than less frequent ones. This dovetails with the usage-based assumption that more frequent patterns in language are entrenched more deeply and are therefore more easily accessible and can thus be retrieved faster, as is evident in e.g. naming tasks. The same mechanism can be

¹That this portrayal of conversion versus derivation is necessarily simplified will become evident in the detailed analysis of corpus evidence in chapter 7.

assumed to apply to natural language production: more frequent chunks are accessed faster, consequently used more often and therefore spread faster. Accordingly, the more frequently a novel pattern is embedded in a very frequent chunk, the faster it can be hypothesized to spread. The importance of chunking for lexical diffusion will be explored for a prototypical case of verb-to-noun conversion in chapter 5. As will be demonstrated, the innovative form can spread fast owing to its embedding in frequently occurring chunks such as the EXISTENTIAL construction *there is a X*.

3.1.3 Processing conversion

It has been claimed above that converted forms are easier to process if inserted into unambiguous contexts. According to Hawkins (2004: 3), grammatical structures result not only from frequency of occurrence (as pointed out above), but crucially depend on “their degree of preference in performance”. By performance, Hawkins (ibid.: 1) means language use and language processing which can be explored in corpora and processing experiments. Preference in performance is a function of complexity and efficiency. Complexity refers to the number of formal units (or “amount of structure”) that needs to be processed (ibid.: 8, 25). Hawkins (ibid.: 8) goes on to explain that “[m]ore structure means, in effect, that more linguistic properties have to be processed in addition to recognizing or producing the words themselves.” Efficient structures, then, are those that have the “lowest overall complexity in on-line processing”, that is, “[t]he (most) efficient structure will [...] be the one that provides the earliest possible access to [...] the [...] proposition to be communicated” (ibid.: 25). Two constraints arise from this: “Express the most with the least” and “Express it earliest” (cf. ibid.). Hawkins (ibid.) formalizes these constraints through three principles: (1) Minimize Domains, (2) Minimize Forms, and (3) Maximize On-line Processing. To explain the potential processing advantage of conversion, the most important of these principles is *Minimize Forms*. It is described by Hawkins (ibid.: 28) as follows.

[I]t is preferable to reduce the number of distinct form-property pairs in a language as much as possible, as long as the intended contextually appropriate meaning can be recovered from reduced linguistic forms with more general meanings [...] by exploiting discourse, real-world knowledge, and *accessible linguistic structure*. [emphasis added]

Accordingly, if the “accessible linguistic structure”, i.e. the context, allows for it, a form should be minimized. In unambiguous contexts, where verb-to-noun conversion can easily be identified as such, conversion is more efficient and therefore easier to process. The

DEVERBAL CONVERTED NOUN construction can consequently be expected to occur in these contexts.

While Hawkins does not distinguish between encoding and decoding (“recognizing or producing”), it is probable that V>N conversion is processed differently by speakers and hearers. Kunter (2015), for example, could show that while in perception hearers generally prefer the synthetic comparative (*friendlier*), speakers (with increasing complexity of the adjective) benefit from producing the analytic comparative (*more friendly*). By analyzing both corpus evidence (production) as well as reaction times and acceptability judgment data (perception), this study endeavors to differentiate between the speaker and the hearer perspective.

3.2 A usage-based account of variety genesis

In his article on “[t]he cognitive evolution of Englishes”, Hoffmann (2014) proposes a way of integrating a Construction Grammar view and the Dynamic Model (cf. Schneider 2007). The combination of a Construction Grammar, i.e. usage-based, approach with the Dynamic Model is a fruitful undertaking considering that most innovations in new varieties happen at the lexis-grammar interface, which is aptly covered by Construction Grammar. It is primarily meso-constructions that are found at this boundary. In the following, the Dynamic Model is explained from a Cognitive Construction Grammar perspective.

In the foundation phase, speakers of two different groups with “different constructional taxonomic networks” interact (Hoffmann 2014: 165). In a process that Hoffmann (ibid.) calls “constructional koinéization”, “infrequent constructions of only a small number of speakers will often be lost, while form-meaning mappings that can be understood by a large number of speakers will become more strongly entrenched”. Among the constructions that are easily understandable are toponyms, which is why toponymic borrowing is frequent. Toponyms can be classified as fully substantive, atomic constructions that occur with a high token frequency since they are “locally salient”.

The second phase, exonormative stabilization, is characterized by borrowing of lexical items that denote flora, fauna, customs, objects or other items of the indigenous culture. These constructions are once again fully substantive and of a high token frequency. Furthermore, they are “found to be peculiar” (Schneider 2007: 39), that is, salient to the settlers.

Nativization, the third phase, is crucial in the development of a new variety. Nativization mostly comes with political independence and thus a greater sense of togetherness of the settler and the indigenous community. On the linguistic front, the result of this increased language contact is the development of a new variety. Most of the linguistic innovations in

new varieties have been located at the lexis-grammar interface (cf. *ibid.*: 83). The explanation for this is readily at hand from a Construction Grammar perspective:

[T]he structural innovations that surface first during Nativization are not completely abstract and schematic macro-constructions. Instead, it is at the meso-constructional level [...] that new and idiosyncratic innovations emerge. From a usage-based Construction Grammar perspective this is actually expected since changes to macro-constructions can normally only occur if they are preceded by significant changes at the subordinate meso-construction level. (Hoffmann 2014: 167)

An increased degree of institutionalization can consequently be interpreted as a “greater use of variety-specific meso-constructions” (Hoffmann 2015).

During the fourth and fifth phase, endonormative stabilization and differentiation, constructions become more complex and also more schematic. This is possible because of an increased availability of meso-constructions, which can then, through the process of abstraction, give rise to macro-constructions. Furthermore, owing to the presence of more abstract constructions, varieties at these phases “can [...] be expected to exhibit a greater type frequency [of certain constructions] (which is more typical of deeply entrenched macro-constructions)” (Hoffmann 2014: 172). Table 3.2 summarizes Hoffmann’s Construction Grammar approach to the Dynamic Model.

The question that arises from Hoffmann’s extension of the Dynamic Model is how the *DEVERBAL CONVERTED NOUN* construction could be integrated. It is an atomic and schematic construction. The alternative, the nominalization via derivation yields a complex and partly schematic, partly substantive construction (if achieved via suffixation; a replaceive form as in *choose* > *choice* can be considered an atomic and substantive construction). Following this account, it can be hypothesized that the atomic construction, conversion, is preferred over the complex construction, derivation, in varieties at the earlier stages. An increase in indigenization should see an increased usage of the more complex construction. However, considering that substrate influence plays an important role in shaping contact varieties, it could also be that conversion persists in the Chinese-substratum varieties, even at more advanced stages, due to extensive transfer from the substratum. This would result in higher type and token frequencies of verb-to-noun conversion in more advanced (Chinese-substratum) varieties.

Table 3.3 summarizes the characteristics of conversion and derivation, with a special focus on their compatibility with the substrata and the lexifier. Even though conversion is more schematic than derivation, it can be expected to be preferred in HKE, the least advanced variety, considering that it offers various advantages: The process is compatible with the Chinese

Table 3.2: Constructions in the Dynamic Model

phase	I	II	III	IV	V
	Foundation	Exonormative Stabilization	Nativization	Endonormative Stabilization	Differentiation
type of construction	fully substantive atomic	substantive-schematic atomic and complex	increasingly schematic increasingly complex		
level	micro-constructional	meso-constructional	meso- and macro-constructional		
frequency	high token frequency	high token frequency	high type frequency		
salience	locally salient (“found to be peculiar”)				
meaning	referential	sets of objects			
examples	toponyms	flora, fauna, culture, customs, objects of indigenous community	phrasal verbs, verb complementation, comparative correlative cxn		

Table 3.3: Conversion vs. derivation in Asian Englishes

	conversion [V] _N	derivation [V + suffix] _N
substantive cxn	?	+
schematic cxn	?	+
atomic cxn	+	-
complex cxn	-	+
cxn compatible with substratum	++ for HKE, SgE, + for IndE	- for HKE, SgE, ++ for IndE
cxn compatible with lexifier	?/+	+
regularization	+	-
redundancy	-	+
ambiguity	+	-
processing effort	+	-

substratum, whereas derivation is not. Furthermore, the DEVERBAL CONVERTED NOUN construction, when embedded in an unambiguous context, regularizes, reduces redundancy and, consequently, also reduces the processing effort. In ambiguous contexts, the processing advantage is lost, so that it is hypothesized that the construction will primarily be encountered in explicitly nominal contexts, e.g. in the NOUN PHRASE construction.

Nevertheless, considering that the DEVERBAL CONVERTED NOUN construction is not fully compatible with the lexifier language, its productivity might be constrained by an increasing degree of indigenization, as witnessed for SgE and IndE. It seems plausible that overriding the constraints on the productivity of contact language features (SYSTEM TRANSFER and LEXIFIER FILTER, cf. section 2.4) will be more difficult and hence more unlikely at later developmental stages. It is therefore hypothesized that the DEVERBAL CONVERTED NOUN construction will be encountered to a lesser extent in the more advanced varieties, SgE and IndE.

A note on diachrony

What Hoffmann's (2014) model illustrates is that constructions can evolve over time and acquire a different status in a variety, in accordance with the development of the variety. A construction will start out as an ad-hoc innovation, "as a feature of an individual mind" (Traugott and Trousdale 2013: 2). Through replication it will eventually be conventionalized and become a new feature of a variety. Traugott and Trousdale (ibid.: 22) call the emergence

3 Theoretical framework

of new constructions *constructionalization*. Conversion is an example of instantaneous lexical constructionalization (cf. Traugott and Trousdale 2013: 30, 189).

According to van Rooy (2011), for such innovative constructs to become actual features of a variety, that is, become conventionalized productive constructions, two aspects are of relevance. The first is “grammatical systematicity” (ibid.: 195). By this, van Rooy (ibid.) refers to a more frequent and more systematic use, i.e. a growing number of systematic attestations in corpora. The second is “acceptability”, which is closely related to systematicity (ibid.: 201). Acceptability can be measured directly, by collecting acceptability judgments, but it can also be measured indirectly, due to the fact that a higher frequency of use by a broader range of speakers must mean that these speakers have accepted the new form. Thus, for a construct resulting from ad-hoc constructionalization to become an innovative construction, to be conventionalized, the construct has to show an expanding systematicity of use and an increasing acceptability. Traugott (2007: 549) summarizes the process of the emergence of an innovative feature as follows:

From a diachronic perspective it is a not unreasonable hypothesis that initially all innovations involve mismatch, in other words, some incongruity of correspondence patterns [...] If speakers adopt an innovative mismatch, by conventionalizing it, they are likely to creatively reanalyze it as a partial match that adds to the repertoire of the language.

“Innovative mismatches” are generally termed ‘errors’, as van Rooy (2011: 192) remarks. However, in contrast to prototypical learner contexts, in New English settings, these ‘errors’ can become conventionalized, thus contributing a new feature to the variety (cf. ibid.: 192–193). Depending on the different paths an innovation might take in different varieties, whether it is reanalyzed or remains a “mismatch”, the outcomes can differ in these varieties. In some varieties, the process of conversion in its entirety or only select micro-constructual conversions might solidify into robust features, yielding new varieties with new local norms.

3.3 Summary

This chapter has laid out the theoretical foundations for the study of verb-to-noun conversion in Asian varieties of English. By adopting a usage-based perspective and analyzing verb-to-noun conversion as a construction in the sense of Cognitive Construction Grammar, it is possible to investigate the emergence of this feature in varieties of English in a way that has hitherto not received enough attention.

Investigating the emergence of verb-to-noun conversions with an eye on frequency as a potential explanatory factor can help predict and explain phenomena such as statistical preemption as well as the faster diffusion of innovative conversions when they are embedded in frequently recurring chunks.

Moreover, this chapter and the previous chapter have charted a usage-based approach to language contact. Two fundamental mechanisms in variety genesis have been set forth: substrate transfer and the degree of institutionalization. The influence of the substrate language(s) on emergent varieties is one of the key mechanisms in language contact, as features that are productive (“compatible”, in Bao’s terminology) in all varieties involved in a contact scenario will be most successful in terms of productivity in the emergent varieties. Productivity is measured by frequency, which explains why it is sensible to assume a usage-based perspective on substrate transfer.

Understanding verb-to-noun conversion as a construction provides a more holistic view of this process than those approaches that have focussed on assigning conversion to either the morphosyntactic or the lexical domain. If V>N conversion is understood as a construction, it means that it is subject to all the constraints and mechanisms that can operate on constructions: The productivity of V>N conversion will depend on the frequency of the competing construction, namely derivation, since constructions giving rise to near-synonyms statistically preempt each other. Furthermore, the processing of V>N conversion will be determined by how well the construction is entrenched, as well as by its degree of schematicity and complexity.

A Construction Grammar approach to institutionalization (cf. Hoffmann 2014) reveals that constructions of different degrees of schematicity and complexity can be expected to be preferred at different developmental stages. Conversion, since it is a morphologically simple process, is assumed to be preferably used at earlier stages. Nonetheless, extensive transfer from an analytic substrate language such as Chinese could lead to V>N conversion appearing even at more advanced stages such as endonormative stabilization.

In the processing of V>N conversion, coercion is an important mechanism. Following Hawkins’s (2004) principles, conversion is expected to pose a processing advantage for the speaker, and potentially also for the hearer. The latter will depend on how fast the hearer can coerce the meaning of the construction. The explicitness of the context is expected to be of particular relevance in this process. The processing advantage of conversion over derivation is hypothesized to play out differently in the varieties investigated, and will depend on how familiar speakers are with this process.

3 Theoretical framework

After describing the theoretical framework for the present study, the following chapter is dedicated to introducing the data as well as the methods used for analyzing them. Considering that previous works on word formation and conversion in varieties of English (cf. e.g. Biermeier 2008; Cannon 1985; Evans 2014) have not produced entirely satisfying results, the present investigation will methodologically deviate from the path taken by these studies and will analyze conversion by drawing on very large corpora (COCA and GloWbE, cf. section 4.1), complemented by experimental methods (cf. section 4.3). Furthermore, the part of the research focussing on non-native varieties of English is only concerned with innovative formations which have not been previously attested in word lists or dictionaries.

Owing to the large size of the data base, it will be possible to investigate in more detail the status of verb-to-noun conversions, distinguishing more clearly between, for example, converted forms that occur in light-verb constructions and full conversions. An important step towards a comprehensive description of verb-to-noun conversion is the overcoming of the traditional dichotomy between grammar and the lexicon, for which Construction Grammar proves to be an effective approach.

The large data base will also help to draw a more detailed picture of how the productivity of verb-to-noun conversion plays out in native and new varieties of English. The aim of this study is not merely to investigate whether the feature exists or not, but rather to give a more nuanced profile of the usage patterns of conversion in distinct varieties (cf. chapter 6) as well as to trace the development of converted forms (cf. chapter 5). This goal can only be pursued by adopting a frequency-based account, situated within the usage-based paradigm, and at the same time keeping an eye on the potential influence which a substrate could have on the productivity of the phenomenon. Domain-general mechanisms of processing are drawn on for further explanation.

4 Data and methods

4.1 Corpora

The introduction of digital technologies has led to a paradigm shift in linguistics that Mukherjee (2009a: 26) goes so far as to call the “corpus revolution”. A corpus is a digital collection of texts. With the possibility of gathering, storing, and searching large amounts of data on a computer, corpus analysis has become a robust method in linguistics. The advantages of corpus analysis are readily at hand. Corpora offer authentic language data, both written and spoken, which supersede the need for introspection and intuitions of individual language users, allowing for generalization. Furthermore, corpora present the opportunity of investigating issues related to usage frequencies, which is highly interesting for work situated in the usage-based paradigm. These frequency data permit researchers to employ statistical methods that have long been used in empirically-based fields of research such as psychology. Statistical methods facilitate generalization and going beyond case studies. Corpora are thus of particular interest in the lexical domain, where dictionaries, i.e. heavily edited texts, had long served as references, e.g. in Cannon’s (1985) study on word formation in US American English. Another advantage of working with corpora is the option of sharing data bases with other researchers. This can promote the replication of experiments and analyses and contribute to the validation of results.

As is always the case, new methods do not only come with advantages, even though these tend to be emphasized at first. While hardly anyone claims that a turn towards empirical research has had negative repercussions on the field, the question of how to carry out such research remains of utmost importance. As far as the method of corpus linguistics goes, the most relevant questions boil down to the following:

representativeness For which varieties (of English) are there corpora available? A better documentation of some varieties e.g. through corpora will lead to a higher visibility of these, potentially at the expense of other varieties.

4 Data and methods

- corpus size** How large should corpora be for a particular task? Increases in size usually result in a trade-off with quality.
- resources** What texts should the corpus contain? Particularly the question of whether the world-wide web should be used as a source of texts has sparked controversial discussions.
- spoken data** Should spoken data be integrated into the corpus? Situational detail that is intricately linked to the production of spoken data is often lost in the transcription process.
- annotation** What are the advantages and drawbacks of annotated data? Is plain text preferable?

These questions are addressed after a presentation of the corpora used in the present study.

4.1.1 *International Corpus of English*

The *International Corpus of English* (ICE) consists of various sub-corpora that seek to represent varieties of English around the world. The project was started in 1990. In January 2016, corpora for the regions of Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, Nigeria, the Philippines, Singapore, Sri Lanka, and the USA are available (cf. The ICE Project 2015). An additional thirteen corpora are in the making (cf. *ibid.*).¹ Each of the sub-corpora contains one million words, of which 60% are spoken and 40% are written language. These one million words come from 500 texts of approximately 2000 words length each (cf. Greenbaum 1996: 5–6). The 32 registers that ICE contains are very diverse; an overview can be found in Greenbaum and Nelson (1996: 13–14). The ICE corpora focus on educated English. All speakers in the corpus are adults “who have received formal education through the medium of English to the completion of secondary school” (Greenbaum 1996: 5–6). Some of the sub-corpora have been tagged and parsed. The Canadian, Hong Kong, Indian, Jamaican, New Zealand, Singaporean and US American sections are available in tagged versions (cf. The ICE Project 2015).

There are several methodological challenges related to the ICE corpora. The first is size. As has been pointed out, the ICE sub-corpora are too small for many language phenomena of low or medium frequency. An investigation of lexical phenomena that is uniquely based on the ICE corpora is likely to be limited by the size of the corpus. Biermeier (2008: 198) is a

¹These include the English language in Australia, the Bahamas, Fiji, Ghana, Gibraltar, Malaysia, Malta, Namibia, Pakistan, Scotland, South Africa, Trinidad and Tobago, and Uganda.

case in point. Another issue is the time that it takes to compile the corpora. The original idea of the founding fathers of ICE was to include only texts from 1990 to 1994 (cf. Greenbaum and Nelson 1996: 5). Some sub-corpora, however, are only being compiled now, over twenty years later. The texts consequently span a long period of time, which can impede comparison. The step-by-step release of the ICE corpora is also due to the fact that the compilation of the corpora according to the strict criteria is not only time-consuming but also costly.

Furthermore, comparison across varieties is complicated by the fact that genres are not global and that some registers are not a part of the local tradition. Greenbaum and Nelson (ibid.) mention that for example “in India class lessons are not dialogues” or that “[i]n Britain and elsewhere, broadcast news is a mixed category – partly scripted monologue (read by the newsreader) and partly public dialogue (brief interviews) or unscripted monologue (statements by public figures)”. This illustrates that applying one scheme to the entire community of World Englishes is almost impossible. These challenges add onto more general issues such as transcription mistakes. Nonetheless, because of its high ‘tidiness’ and meticulous compilation (including metadata on the speakers), ICE is still *the* reference corpus for World Englishes research.

4.1.2 *Corpus of Contemporary American English*

The *Corpus of Contemporary American English* (COCA, Davies 2008–) is a monitor corpus of the US American variety of English, i.e. data are continuously added to the corpus. COCA was released in early 2008 (cf. Davies 2009: 159) and is available online. Searches can be executed via a web interface that all corpora compiled by Mark Davies and team share. A detailed description of the corpus and its application to the study of English is available in Davies (ibid.) and Davies (2010). In what follows, only the most important aspects of the corpus shall be presented.

COCA was compiled, firstly, as a monitor corpus of a native variety of English. Secondly, it was to be the first large corpus on the American variety. At the time, the *American National Corpus*, modeled on the *British National Corpus*, had not been completed (cf. Davies 2009: 159–160).² COCA is thus a first in various respects: it is the first large corpus of American English, “the first reliable monitor corpus of [the] English [language]” (Davies 2010: 447), and it is (at least in parts) freely available to a large research community via the internet.

²The current, second release of the ANC dates from 2005 and contains over 22 million words. However, this second release is not balanced. The compilers are aiming for a final release that comprises 100 million words, but a lack of funding opportunities seems to slow down the process (cf. American National Corpus Project 2012a,b).

4 Data and methods

What is meant by “reliable” is that in contrast to other large corpora of English that cover a considerable time span (such as the *Bank of English* or the *Oxford English Corpus*), COCA offers a balance of genres and also sub-genres. For every year, the percentage of words coming from the diverse sub-genres remains constant (cf. Davies 2010: 453). This facilitates a comparison of language phenomena across registers over a long period of time.

For every year from 1990 on, COCA contains around twenty million words (cf. Davies 2009: 160). At the time of conducting the corpus analyses for the present study, 2012 constituted the latest year represented in the corpus. For 2012, only half the amount of data was available (11,363,451 tokens), which is why the year 2012 is not always taken into consideration in the following analyses. In December 2015, COCA was updated to include data from the years 1990 to 2015, so that in its present state (November 2016), the COCA corpus consists of approximately 530 million words.

The genres that COCA comprises are spoken language, fiction, popular magazines, newspapers, and academic journals. They include the following types of texts (ibid.: 161–162).

SPOK	“Transcripts of unscripted conversation from more than 150 different TV and radio programs”
FIC	“Short stories and plays from literary magazines, children’s magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts”
MAG	“Nearly 100 different magazines”
NEWS	“Ten newspapers from across the US”
ACAD	“Nearly 100 different peer-reviewed journals”

Each of these registers represents about a fifth of the corpus (cf. ibid.: 160) and is composed of various sub-genres such as academic texts from the field of education or philosophy or newspaper articles from the sports and the financial section. The corpus does not sample any texts from the internet; the various reasons for this decision can be found in Davies (ibid.: 162–163).

It is immediately evident that classifying texts by their source (newspaper, magazine etc.) cannot be a very accurate way of doing so since it disregards text-internal characteristics. Focussing only on the medium of publication as the distinguishing criterion may lead to overlaps of the proposed genres. An article from the newspaper section of the corpus that appeared in the financial section of a newspaper is likely to be fairly similar to an article published in a magazine that deals with financial topics. Furthermore, the magazine article

about finances is highly likely to share fewer characteristics with an article about gardening or home decor than with the newspaper article about finances, even though the two magazine articles share the publishing medium and would consequently be subsumed under the same heading in COCA. The notion of *genre* as it is used in COCA is thus not the highly technical notion that is applied in text linguistics (cf. e.g. Biber 1988). It is rather a rough characterization of the texts contained within that section of the corpus. In the following, for reasons of practicality, the term *genre*—whenever used in the context of COCA—is meant to refer to the genres as given in COCA, not to the text linguistic notion.

Even though the entire corpus is tagged according to the CLAWS7 tag set (cf. University Centre for Computer Corpus Research on Language 2015), the CLAWS tagger utterly fails when dealing with conversion. When searching for *disconnect*, COCA returns 1165 hits. When searching for *disconnect* as a noun³, COCA returns 0 hits, even though manual POS tagging as done in the study presented in chapter 5 reveals that there are 743 instances of *disconnect* being used as a noun. Due to this discrepancy, in this study, the tagging that comes with COCA is disregarded (unless indicated) and all tagging is performed either manually or by a custom-made computer script.

4.1.3 Corpus of Global Web-based English

The *Corpus of Global Web-based English* (GloWbE, Davies 2013) is a web-based corpus of twenty varieties of English released in 2013. It is available via the internet and can be searched with the same interface as COCA. The aims that have led to the creation of GloWbE are threefold (cf. Davies and Fuchs 2015b: 3–5). The first consideration was that ICE, the corpus that is broadly used for studies on varieties of English, is too small to give representative data of low- and medium-frequency language phenomena (cf. *ibid.*: 2). In order to analyze these phenomena, a larger corpus is needed. Secondly, this new corpus was devised to be comparable to ICE as far as “genre balance” is concerned, i.e. to be composed of roughly 60% of spoken/conceptually oral data (*ibid.*: 3–4). Davies and Fuchs (*ibid.*: 4) consider blogs to contain near-spoken data and thus included 60% of blogs and 40% of other web pages in the corpus. (Cf. section 4.1.4 for a discussion of the genre of websites.) Thirdly, the corpus should represent different varieties of English, as the ICE corpora do (cf. *ibid.*: 2–3). GloWbE was then compiled out of 1.8 million web pages (cf. *ibid.*: 5). For the exact compilation procedure the reader is referred to Davies and Fuchs (*ibid.*: 3–5). GloWbE comprises 1.9 billion words

³disconnect.[nn*]

and is therefore over 150 times larger than all ICE sub-corpora taken together (cf. Davies and Fuchs 2015b: 25). Table 4.1 lists the sizes of the GloWbE sub-corpora used in this study.⁴

Table 4.1: Relevant GloWbE sub-corpora by size

Great Britain	387,615,074
United States	386,809,355
India	96,430,888
Singapore	42,974,705
Hong Kong	40,450,291

Despite the fact that Davies and Fuchs (*ibid.*) present promising case studies, it is necessary to address some unresolved issues. The first lies in the compilation of the corpus. In determining where the language on web pages is from, Davies and Fuchs (*ibid.*: 4–5) rely on Google. For web pages with a country top-level domain (e.g. .sg for Singapore or .hk for Hong Kong), Google assumes that these sites are hosted in the corresponding countries. For other web pages, the algorithm programmed by Google takes into account the IP address of the web server in question, the links to that website and the visitors of that website (cf. *ibid.*: 4). This procedure is problematic in some respects. The first is that Google does not disclose their algorithms so that it is impossible to know how Google exactly determines from what countries web pages are. Despite the opacity of this process, Davies and Fuchs (*ibid.*: 5) claim that they “have yet to find a single website whose country has not been correctly identified by Google”. The second challenge is that speakers from other countries may host sites with a certain country top-level domain; a German, for example, could buy a domain ending in .hk without difficulty. Furthermore, speakers from other countries can easily contribute to pages that are not hosted in their own country. Speakers of Singapore English, for example, could go to the website of any British newspaper and post in the comments section—and the other way around. Although Davies and Fuchs (*ibid.*: 26) acknowledge that this can happen, they do not address the question further. Cook and Hirst (2012: 281), however, show that “English Web corpora from national top-level domains may indeed represent nation dialects”, so that the procedure adopted for the compilation of GloWbE can be accepted as a viable method. The question that remains is whether corpora obtained on the basis of different top-level domains are comparable (e.g. whether they sample the same number of websites belonging to a particular genre such as newspaper articles, cf. *ibid.*: 291).

Another issue, which is a general characteristic of web-based corpora, is non-standard orthography. “[T]he web typically values content creation above perfection and tolerates ill-

⁴The sub-corpora are labelled like the countries where the websites constituting the corpora are hosted.

formed language” (Fletcher 2007: 36), which is why the web is particularly prone to spelling mistakes. In GloWbE, evidence for this is not hard to find: *b4* for *before* occurs 3097 times, ’s with a plural instead of a possessive meaning as in *American’s* is frequent, and there are no attempts at standardizing orthography as in the compilation of the ICE corpora. This consequently leads to blatant problems when it comes to tagging the corpus. As Mair (2015: 29–30) points out, a high rate of inaccurate tagging is particularly obvious in varieties such as Nigerian or Jamaican English where, in informal language, pidgins and creoles are mixed with English. As an example, Mair (ibid.: 30) mentions *inna*, the Jamaican creole variant for *in*, which is “generally and mistakenly tagged as a noun”.

Finally, another aspect that is worth considering is the quality of the texts sampled. In the careful compilation of for example the ICE corpora every single text is read by a researcher. This is impossible for a corpus of the size of GloWbE. Consequently, GloWbE is likely to contain nonce-texts. The following are examples of hardly intelligible language.

- (4.1) Family vacations accept acquired over the years. There accept never been added choices, added array and added options accessible for ancestors vacations as there is today. (GloWbE-HK, G⁵)
- (4.2) According to another Korea media reports, the action of electromagnetic steel belongs to cater to environmental protection era of sell like hot cakes steel varieties (GloWbE-HK, G)
- (4.3) North Branch building group of the carry flag Qingdao blue biomedical industry park (GloWbE-HK, B)

For these corpus samples, it is far from trivial to determine whether one is dealing with non-standard language features or computer-generated spam that cannot be regarded as ‘real’ language. However, it is nearly impossible to double-check. Theoretically, it is possible to trace the source pages of the corpus, but in most cases the web pages in question do not exist any more. Yet, for example 4.3 the source page⁶ still existed when this project was started. Reading the top part of the website, one finds a fairly coherent text in standard English. Nonetheless, scrolling down to the comments section one notices lots of spam, i.e. computer-generated entries, among it also one in Cyrillic script. Example 4.3 stems from this comments section and can therefore hardly be considered an example of Hong Kong English.

⁵G indicates that the text stems from the general section of the corpus, B stands for the blog section.

⁶<http://www.beautyandhealthreviews.com/think-you-know-all-the-beauty-tips-try-these/comment-page-51/>

Despite these issues, GloWbE is still a useful resource in the study of World Englishes. It is very large and the sheer amount of data seems to compensate for other methodological weaknesses. Particularly when it comes to lexical phenomena, GloWbE, because of its size, is an unprecedented resource for research into World Englishes, especially for near-spoken language. At the moment, GloWbE is one of the best data sources for projects on lexical variation in World Englishes; it is free, fast, and vast. (An evaluation of GloWbE as a resource for World English studies is provided in the discussion in chapter 9).

4.1.4 Potential and limitations of corpora

After presenting the corpora that are used for the present studies, it is necessary to address the questions that have been raised above. They are repeated here for convenience.

- What makes a corpus representative? What varieties are documented and how does that influence the researcher community?
- How large should corpora be?
- Should the world-wide web be used as a source of texts?
- Should the corpus contain spoken material?
- Should the corpus be annotated?

Representativeness

A crucial question in the compilation of corpora is representativeness. There is no unanimous agreement on the definition of representativeness (cf. McEnery and Hardie 2012: 10), but extensive discussions of the notion are provided in Biber (1993) and Leech (2007). Here, suffice it to say that representativeness is generally understood as referring “to the extent to which a sample includes the full range of variability in a population” (Biber 1993: 243). Population does not necessarily have to refer to speakers but can also refer to the population of genres or text types. During corpus compilation, compilers should strive for representativeness, even though some have claimed that this goal is unattainable (cf. Váradi 2001: 588).

A study of World Englishes, however, has to go beyond the question of how to exactly define and also achieve representativeness. In studying varieties of English, it is indispensable to take a broader approach by asking what varieties of English are documented (for linguistic purposes) and whether the existing documentation is representative of the varieties of English, since documentation influences research.

In order to do so, the concept of *doculects* as introduced by Cysouw and Good (2013) and Cysouw (2014) is drawn on. A doculect, according to Cysouw and Good (2013: 338), is the basic entity in the definition of a language and also of language studies. A doculect is any documented lect. The motivation for reconsidering what makes a language is that

while what constitutes a specific language in some abstract sense will perhaps always be controversial, there is no controversy in simply saying that there exists a given book, sound file, manuscript, or article that contains data documenting some language variety, even if there is disagreement about how that variety should be classified. (Cysouw 2014)

Consequently, the existence of documentation is the basis for the existence of the lect, i.e. language documentation makes language. As Cysouw and Good (2013: 338, 344) say, “[e]ven a language only known by name would count as documented in the present context”, since “the minimum requirement for making a language variety ‘real’, at least for the linguist, is the pairing of a resource with a glossonym⁷”. This approach can thus be seen as a “reversal of the code-source relationship” (Cysouw 2014) in that documentation precedes language. As a consequence, languages that are not documented do not ‘exist’ because they are not visible to the (researcher) community.

Hence, in the study of varieties of English, one has to keep in mind that some varieties are better documented than others and therefore more visible than those varieties that have not been as minutely documented. (Researchers’) perceptions of varieties of English will be shaped by what resources are already available in the community. That is, varieties that are already well documented will also be the object of further study. Or, as Leech (2007: 134) asserts, “research is skewed by what resources we can lay our hands on”.

An example of this is the International Corpus of English, one of the major resources in English variationist research. At this point, ICE contains five Asian varieties of English and two African varieties. As a result of this documenting practice, there is a large community of researchers who work on Asian varieties. African varieties, however, probably also because they have not been part of ICE to the same extent, have not received an equal amount of attention.⁸ The tide seems to have been turning over the last years, though, with four new sub-corpora of African varieties in the making (Ghana, Namibia, South Africa, Uganda).

Moreover, it is of crucial relevance to consider the representativeness of corpora available for individual varieties. The mere existence of a corpus does not necessarily make for a

⁷By *glossonyms* Cysouw and Good (2013: 339–340) understand the labels that are used to refer to languages.

⁸A cursory glance at two major journals in the field confirms this impression. There is, for example, a special issue of the journal *English World-Wide* on Asian Englishes (Vol. 30, No. 2, 2009) but not on African Englishes. The same is true for *World Englishes*, that featured special issues on Singapore English in 2014 (Vol. 33, No. 3) and on English in China in 2015 (Vol. 34, No. 2) but has not published anything similar on African Englishes.

truthful representation of actual language use. Depending on the available resources, the doculect need not (and in many cases probably does not) reflect the full spectrum of language use. The ICE corpora can once again serve as an example. While bilingualism is a reality for most speakers in ESL countries, ICE does not render speech that is produced in a language other than English. This makes it more difficult for researchers to assess, for example, how and why English is combined with other languages (code switching).

Academic discourse is consequently very much subject to prevailing documentation practices. While it would be desirable to free research from the constraints of documentation and the reliance on doculects, this objective is of course utopian. Corpora, as a key instrument of linguistic research, must therefore always be seen in light of the fact that what is documented cannot be assumed to match all dimensions of actual language use.

Small vs. large corpora

This section aims to critically reflect on the benefits of small and large corpora. ICE comprises one million words per variety, whereas COCA and GloWbE comprise 450 million and 1.9 billion words respectively. With 13 sub-corpora of size one million words each, ICE is thus a comparatively small corpus of the English language, whereas COCA and GloWbE offer vast amounts of data. While size can be a limiting factor—and has proven to be, particularly in the domain of word formation (cf. e.g. Biermeier 2008)—it can also present an opportunity. Generally, size and tidiness of corpora can be seen as antagonistic goals in corpus compilation (cf. Davies and Fuchs 2015b: 26, 2015a: 47). With an increase in size comes a lower degree of tidiness, while smaller sizes allow for a more careful revision of the data. Corpus compilers therefore have to decide which of the two, size or tidiness, their primordial goal is and then act accordingly. The ICE compilers have opted for tidiness; in contrast, the corpora compiled by Davies and team are large in size but also less neat. What this means for corpus analysis is laid out in the following.

Smaller corpora facilitate annotation, simply because they make the task of manually correcting the work of automatic parsers and taggers more feasible. Small corpora, especially the ICE corpora, are very ‘tidy’ also in the sense that the texts that the corpora contain have deliberately been chosen by the researcher as worth of finding their way into the corpus. This is not only reflected in the careful annotation of orthographic mistakes in ICE, but also in the compilation of texts of many different registers (for a complete list of registers that the ICE corpora contain cf. The ICE Project 2009) as well as the inclusion of meticulously collected metadata (e.g. speakers’ age, languages, education, occupation in ICE-Canada, cf.

Columbus 2010).⁹ This high degree of control is not possible for extremely large corpora such as GloWbE, which is consequently ‘messier’. There are no spelling checks applied to the data and annotation is limited to part-of-speech tagging by an automatic tagging software. Furthermore, GloWbE does not contain any metadata on the authors of the texts. Also, the quality of the texts in GloWbE is not as uniform as in ICE. GloWbE represents a special case in this respect as it is the largest web-derived corpus at this point. It is impossible to read all texts that find their way into the corpus, which inevitably leads to the accidental inclusion of nonce-texts such as computer-generated spam (cf. example 4.3 and below in the discussion of web-corpora).

The careful selection, classification and compilation of a balanced corpus such as the ICE sub-corpora consumes large amounts of time. While small corpora have the big advantage of tidiness, a major drawback is that the time required to compile them can be so long that the data are almost outdated when the corpus becomes available. COCA, on the other hand, as a large monitor corpus, comes with a smaller number and less careful balancing of registers and less precise tagging, but has been updated regularly since its first compilation in 2008. The conclusion to be drawn from this comparison of small and large corpora is that in a thorough linguistic analysis, it is adamant to combine both small and large corpora.

Web-based corpora

While many researchers acknowledge the potential of the web as a vast source of linguistic information (cf. e.g. Fletcher 2007: 27; Mair 2007: 235; Mukherjee and Schilk 2012: 197–198),

[t]he main problems with the first approach [= web as corpus] are that we still know very little about the size of this ‘corpus’, the text types it contains, the quality of the material included or the amount of repetitive ‘junk’ that it ‘samples’. Furthermore, due to the ephemeral nature of the web, replicability of the results is impossible. (Hundt et al. 2007: 2–3)

Not much has changed since Hundt et al. (ibid.) wrote this in 2007. The web remains as elusive as it was. Nonetheless, it is the only source that provides informal language from all over the globe so fast and easily. Therefore, in their attempt to create a very large corpus of informal language of World Englishes, Davies and Fuchs (2015b) considered the web to be the ideal source. This resulted in a corpus that is over 150 times larger than all ICE sub-corpora taken together (cf. ibid.: 25).

⁹Nonetheless, it has to be noted that the reliability of the metadata in ICE varies between the subcorpora due to the fact that there are no uniform guidelines on how to handle the metadata. ICE-India, for example, shows comparatively inaccurate metadata (cf. Hansen 2015).

Many of the challenges that web-based corpora present have already been addressed above in the section on the *Corpus of Global Web-based English*. Nonetheless, many of these are compensated for by the large size that web corpora generally have. The key, once again, is to combine several types of corpora. As Hundt et al. (2007: 4) point out,

studies also show that—despite the many unsolved methodological problems—web data can provide useful additional evidence for a broad range of research questions, especially if combined with results from standard reference corpora.

Other corpus linguists like Hoffmann (2009: 37), Mair (2007), and Mukherjee and Schilk (2012: 191) also insist on a combination of established corpora such as ICE and new, web-based corpora. This is achieved in combining GloWbE and ICE, two corpora that, according to Davies and Fuchs (2015b: 26), “complement each other nicely”. In a recent study, Heller and Röthlisberger (2015) were able to show that as far as the dative and genitive alternation in English go, ICE and GloWbE do not offer significantly different results. In logistic regression models estimating the odds of either alternation they found that regardless of which corpus constituted the data base, the importance of almost all predictors for the estimates of the regression model remained the same.

When working with web-based corpora, one necessarily has to answer the question of register. “What the precise relationship is between informal digital literacy and actual spoken language is an extremely tricky issue”, as Mair (2015: 30–31) points out. In their compilation of GloWbE, Davies and Fuchs’s (2015b: 4) intent was to emulate the 60% to 40% relationship for spoken to written genres that is the basis of the ICE corpora. In order to do so, 60% of GloWbE is made up of web pages containing blogs, since Davies and Fuchs (*ibid.*: 26) consider this type of texts to come closest to spoken language. In their description of the corpus (*cf. ibid.*), they do not elaborate on this conclusion any further. That the classification of text types on the web is not accomplished easily is demonstrated by e.g. Kailuweit (2009): Many attempts at expanding the well-known orality-literacy continuum by Koch and Oesterreicher (1985, 2007) have been made, many of them failing to grasp the text types of the web in their entire complexity. This is mainly due to the fact that on the web, the well-established “clear-cut distinctions between spoken and written language” are “blurr[ed]” (Gatto 2014: 51), as text types such as chats or tweets (text messages posted on Twitter) show. In one such attempt to classify what can be found on the web, Biber and Kurjian (2007) identify eight different clusters (i.e. text types) with the help of a multi-dimensional analysis of the web. These clusters differ on four different dimensions of variation, namely personal or involved narration, persuasive or argumentative discourse, addressee-focused discourse, and abstract or technical discourse (*cf. ibid.*: 116).

In a study of a random sample of the general section of GloWbE-US, GB, CA, AU, NZ, Biber et al. (2015: 24) show that the main communicative purposes on the web are narration (31%), informational description/explanation (14%), and opinion (11%). They further find that on the web there are many texts (30%) that combine two registers, i.e. that follow more than one communicative purpose. Generally, there is a large quantity of “specialised web registers not found in print media” such as discussion forums (ibid.: 29).

Tagliamonte (2013 in Tagliamonte 2014: 229) finds that speakers differ in their use of language depending on the type of digital text that they are producing (e-mail, instant messaging on computers, texting on phones) and the device that is used for the production of these texts. In analyzing texts from the internet it is therefore necessary to identify the text type that one is dealing with and decide for each case whether it is rather on the oral or the written end of the continuum. Only after such a careful analysis is it possible to identify potentially similar ‘traditional’ registers. However, it is highly likely that in some cases there are no corresponding ‘traditional’ registers for text types on the web. As Crystal (2011: 21) states:

Internet language is identical to neither speech nor writing, but selectively and adaptively displays properties of both. It is more than an aggregate of spoken and written features. It does things that neither of the other mediums does.

The web thus not only displays greater variation in genres in general (cf. Rowley-Jolivet 2012: 147), but also seems to be the home of unprecedented genres such as blogs that have emerged because of the diverse communicative opportunities that have arisen with the medium of the internet (cf. Garzone 2012: 237). When it comes to GloWbE, the coarse classification of texts from the web into a ‘general’ and a ‘blog’ genre as representing near-written and near-spoken language should therefore be taken with a grain of salt.

As far as the blog register, “the quintessential genre of the searchable web” (Biber et al. 2015: 40) and also the main component of GloWbE, is concerned, Mair (2015: 31) questions “whether blogs constitute a recognisable genre”. The findings by Biber et al. (2015: 40) indicate that blogs “vary widely in their situational characteristics and communicative purposes”. Thus, the existence of a unified ‘blog’ genre has to be rejected. In their classification of different blog text types, Grieve et al. (2010: 303) identify two major types of blogs, one concerned with personal topics, comparable to a diary, and one informational, in which authors comment on different topics. In general, they find that blogs constitute a distinctive text type that shows a peculiarity that is unprecedented in other studies on textual variation (cf. Biber 1989; Biber and Kurjian 2007). Garzone (2012: 237) goes so far as to “consider the blog as a macrogenre” in itself. A main difference between blogs and other text types is that for the for-

mer, Grieve et al. (2010: 309–310) find—via a factor analysis—that the personal dimension of variation does not include second person pronouns. That is, texts that score high on the personal dimension (mostly blogs of the personal diary blog type) “have one major topic: their author”. In other text types, personal involvement usually comes with markers of speaker-hearer (or writer-reader) interaction such as second person pronouns. This is not the case for blogs, which consequently makes them a text type worth analyzing independently. While both blog types identified by Grieve et al. (*ibid.*: 320) make use of a comparatively “personal and conversational style”, they do not consistently offer features that are also found in spoken speech. This is particularly the case for the commentary blog type, that is among others characterized by a fairly nominal style, associated with informational density (*cf. ibid.*: 317). It thus seems that while blogs have emerged as a stable text type in the dynamic environment (*cf. Santini et al. 2010: 13*) of web genres, they do not represent spoken language use. Any comparison between GloWbE and other spoken corpora such as the relevant sections of the ICE corpora should therefore be interpreted with this difference in mind.

In what follows, I will thus assume that texts on the web can generally be thought of as less formal than written (printed) text but also as less informal than spoken language. This is based on two considerations. First, the mere act of typing a message is expected to moderate language production and lead the text away from the spoken end of the continuum. Nonetheless, as Fletcher (2007: 36) points out, on the web very often the speaker’s main aim is to achieve their communicative goal at the expense of stylistic considerations or grammatical correctness. This fact is believed to lead a text away from the written end of the continuum, with texts on the web consequently occupying the middle ground between spoken and written language.

Treatment of spoken data in corpora

Spoken data are extremely rich in nature. They do not only comprise what is said but also how it is said. That is, the phonetic layer in itself is full of intricate detail as it includes phones but also suprasegmental features. Furthermore, speech is also characterized by its situatedness. Situational detail includes paralinguistic cues and other non-verbal features such as gaze, gestures, facial expressions and many more. Very often, because of the complexity of the data, situational detail is not transcribed (*cf. McEnery and Hardie 2012: 4*), so that what finds its way into the corpus is merely a representation of what was said. A large part of what makes speech can therefore be said to be ‘lost in transcription’. Phonetic features are usually only included in corpora that have been compiled for specifically phonetic research purposes. The ICE corpora, for example, mainly serve to investigate morphosyntactic variation in World

Englishes. Therefore, they only contain orthographically standardized transcripts of spoken data that can hardly be considered “a reliable source of evidence for research into variation in pronunciation” (ibid.). A further problem related to transcription are spoken data that have been transcribed by non-trained non-linguists. Non-linguists might underestimate the relevance of some features of spoken speech and omit or change some of the original wording as they transcribe (cf. ibid.), which necessarily results in a distortion of the original data.

Despite the challenges that the compilation of spoken corpora poses, spoken data are in fact highly valuable to linguists interested in linguistic innovation. Innovations are most likely to first occur in spontaneous speech and then, if at all, make their way to the written registers. A corpus that is exclusively based on written texts will therefore unavoidably fail to capture the most innovative features of a variety. On the other hand, as has been pointed out above, corpora of spoken material are a lot more costly to compile, so that mega-corpora such as COCA or GloWbE only include written texts or transcripts (e.g. of broadcasts) that are readily available.

In the case of verb-to-noun conversion, a phenomenon that is very rare in the native varieties of English, the use of large corpora is indispensable, which comes at the expense of phonetic information. Therefore, in this study, corpus size trumps phonetics, which means that all aspects related to the phonological side of conversion, e.g. stress shift as in *to tormént* - *a tórmént* (cf. Plag 2003: 110), will not be targeted in the analysis.¹⁰ Nonetheless, complementing data from the mega-corpora with data from the ICE corpora, particularly from the spoken section, it becomes possible to at least analyze conversion and other phenomena in their discourse-pragmatic context in unscripted interaction.

Corpus annotation

A further point that is of interest for corpus analysis is the question of whether plain text or annotated text is preferable for linguistic analysis. Nowadays, many corpora are annotated automatically by special software such as the CLAWS tagger (University Centre for Computer Corpus Research on Language 2015; cf. Gatto 2014: 17). Annotation by parts of speech (tagging) is fairly common, while syntactic parsing is comparatively rare because of its complexity. Annotation generally facilitates searches and also enables the researcher to perform more complex searches, which is particularly useful for very large corpora.

¹⁰Another reason to disregard the phonological side of conversion is the fact that there is no consensus as to the status of pairs such as the one mentioned above. Plag (2003: 110), for example, does not consider these to be “clear cases of conversion, because the relationship between the pairs is marked overtly, even though this marking is done not by an affix, but by a prosodic property.”

Nevertheless, automatic annotation is highly error-prone particularly when dealing with non-standard or spoken language data. Orthographically incorrect spelling or missing or incorrect punctuation can gravely impact on the accuracy of part-of-speech tagging (as Mair 2015: 30 points out for the Jamaican section of GloWbE). This issue is aggravated when it comes to spoken data. The fact that speech is unplanned and therefore shows false starts, repairs, anacolutha and the like makes spoken data too complex to be processed automatically. In addition, since word stress is usually not transcribed, part-of-speech tagging is particularly unreliable for homographs that occur in ambiguous contexts.

Most automatic tagging is thus not accurate enough to *not* be post-edited by a human being. However, manual tagging is quite laborious and almost impossible for corpora of the size of COCA and GloWbE. In many cases, the tagging that comes with the corpora is consequently not accurate enough to rely on it. If parts of speech are not identified correctly, searching for parts of speech will actually result in highly distorted results instead of facilitating or improving the analysis (see the example of *disconnect* in COCA mentioned above). Due to these “issues of accuracy and consistency” in tagging, the “purity” of the untagged data is sometimes more appealing to corpus analysts (cf. McEnery and Hardie 2012: 14). As has been pointed out, part-of-speech tagging is extremely unreliable in the case of conversion, so that in this study, unless otherwise indicated, tagging is carried out by a custom-made, conservative computer script, and all remaining parts of speech that the script could not identify are then annotated manually.

4.2 Quantitative methods

In order to make large amounts of linguistic observations accessible, corpus research usually goes hand in hand with statistical methods of analysis. The main quantitative methods applied in the following studies will be presented in this section. Among them are collocation analysis, and linear and logistic regression. They are all methods that try to establish a relation between previously specified independent and dependent variables based on the data points that result from observation. In corpus linguistics, these methods are often denominated corpus-based methods. They contrast with corpus-driven approaches where variables of analysis are not specified *a priori* but “emerge” in the course of the analysis (Biber 2010: 162).

4.2.1 Collocation analysis

“[C]ollocation is the statistical tendency of words to co-occur” (Hunston 2002: 12). The analysis of these co-occurrence tendencies of words is called collocation analysis. One of the most popular measures of collocational strength is the point-wise mutual information score (MI-score). While there are many other association measures,¹¹ the mutual information score is used to calculate collocational strength in the corpora by Davies and team (cf. Davies n.d.).

The MI score indicates the “mutual dependence of [...] two words” and is generally calculated as follows (Metin and Karaođlan 2011: 177):

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x) \cdot P(y)} \quad (4.4)$$

$P(x)$ and $P(y)$ represent the probabilities of occurrence of two potential collocates x and y and $P(x, y)$ is the probability of these two words “coming together in the text” (ibid.). In COCA and GloWbE, “coming together” is operationalized with the help of the span of words, that is, through indicating how close to x y must occur. The corresponding measure is the number of words to the left and right of the node word. In the Davies corpora, the MI score is thus calculated as follows (cf. Davies n.d.):

$$\text{MI} = \log_2 \frac{F(xy) \cdot \text{corpus size}}{F(x) \cdot F(y) \cdot \text{span}} \quad (4.5)$$

$F(x)$ is the token frequency of the node word, $F(y)$ the frequency of the collocate. $F(x, y)$ is the number of times that y appears in proximity to x , whereby proximity means within the range as specified by the span. The span is calculated by adding the number of words allowed to either side of the node word. The size of the corpus is included in the equation so as to yield probabilities. In short, what is calculated is “[a] comparison of observed and expected frequencies of pairs of words” (Stubbs 1995: 31). This method is based on “the assumption that it is meaningful to compare (a) a real corpus and (b) a hypothetical corpus consisting of the same words in random order” (ibid.). The MI-score is then a “significant deviation[.] from hypothetical randomness” (ibid.).

The weaknesses of the MI score have been pointed out by various scholars, among them Gries and Mukherjee (2010). One main disadvantage of the MI score is that its calculation is based on the assumption that all words are equally likely to cooccur, that is, are equally independent of each other. This, as Gries and Mukherjee (ibid.: 523) point out, is “almost never the case in natural language”. They argue that, for example, “the probability of *of* two words after *in* is very much higher when the word immediately after *in* is *spite*”. This means

¹¹Evert (2004) lists a great variety of them, Gries and Mukherjee (2010) and Gries (2013a) propose even more.

that “associations are not necessarily reciprocal in strength” (Ellis and Ferreira-Junior 2009: 198). While *spite* might increase the probability for preceding *in* and succeeding *of*, neither *in* nor *of* would intuitively be expected to collocate particularly strongly with *spite*. Due to this flaw in the calculation of the MI score, words that are very infrequent but occur very frequently in close proximity to a certain node word will obtain an extremely high MI score (cf. Mukherjee 2009a: 104). Consequently, the absolute MI score values should be taken with a grain of salt, bearing in mind that particularly high numbers might be due to skewed distributions in the corpus. A further drawback of the MI score is that its calculation requires a fixed span of words preceding and following the node. This makes it impossible to allow for varying lengths of collocations.

Nonetheless, the MI score is a viable tool to analyze co-occurrence probabilities. The problem of very high scores for infrequent lexemes is only acute for very idiomatic, very substantive constructions such as *in spite of* or proper nouns. Yet, these can easily be identified as such by a qualitative analysis of the corpus tokens. In cases where only the part-of-speech of the collocate is of relevance, e.g. in order to determine the part-of-speech of the node word, this peculiarity of the MI score does not pose a problem.

4.2.2 Linear regression

Linear regression¹² is a method to establish a linear relation between variables, such as year and frequency. One of the variables depends on the other(s), in this example the hypothesis could be that frequency varies depending on the year. In a linear regression, a line is fitted to the data so that “the line is as close as possible” to every single data point (Baayen 2008: 85). The method with which this is achieved is called least squares regression. The idea is to “minimiz[e] the squared vertical differences between the data points and the line” (ibid.: 86). By reducing the distances between the fitted line and the data points the line is approximated to the points, resulting in a line that best describes the relation between the data points.¹³

While this method is problematic for skewed data—few outliers could ‘push’ or ‘pull’ the line in another direction, i.e. make it steeper or more gentle—it is suitable for evenly distributed data (cf. ibid.: 92). In chapter 5, linear regression is applied to data from the monitor corpus COCA. The data include observations for every single year out of 22 years so that data skewness is not an issue.

¹²Linear and logistic regressions were calculated with the open source programming language R (R Core Team 2014) using the integrated development environment RStudio (RStudio n.d.).

¹³Linear regressions were fitted with the `lm()` command in R.

Interactions

In the above-mentioned example, only two variables, time and frequency, are considered. When calculating statistical models it is also possible to specify more than one independent variable and also interactions between variables. An interaction between variables is illustrated in the two graphs in figure 4.1.

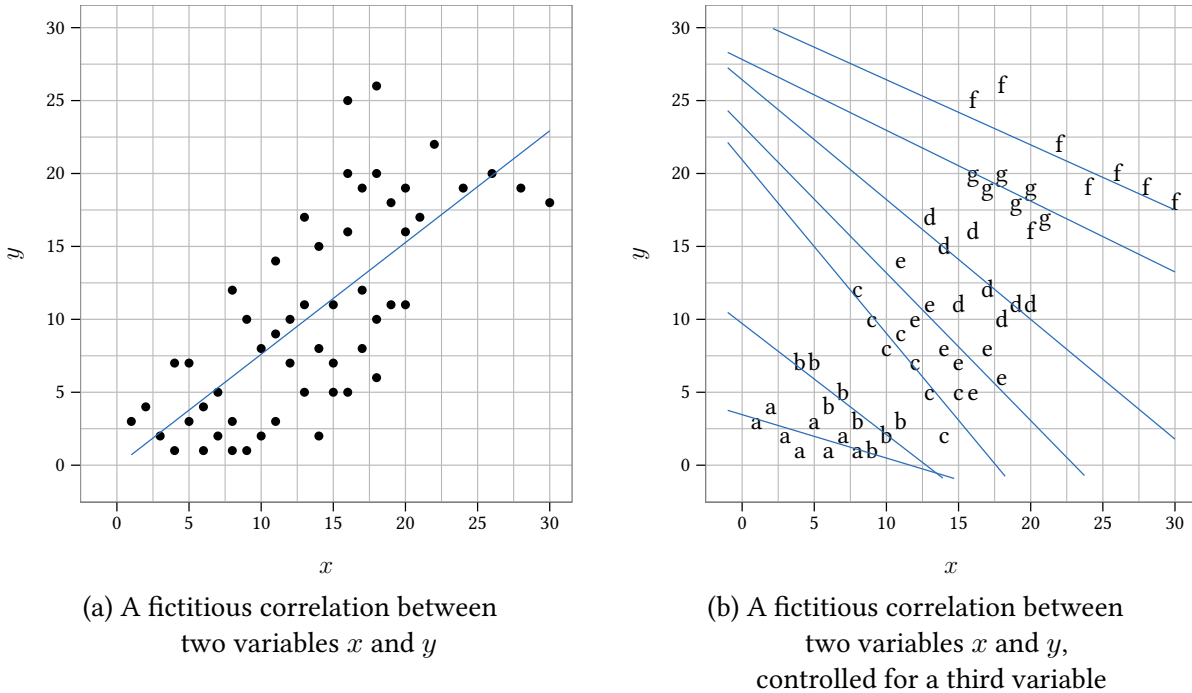


Figure 4.1: Interactions between variables, taken from Gries (2009: 6)

What can be observed here is that a relationship between two variables x and y might change when taking into account a third, moderating variable (also called covariate). In the case of time and frequency, one could assume that another factor, for instance variety (exemplified by the letters in figure 4.1b), influences how frequency changes over time. In other words, if in a study the frequency of a linguistic phenomenon is expected to depend on the development of individual varieties over time, it is convenient to specify interactions in the corresponding model.¹⁴

Interpreting the output of a linear regression

The output of a linear model calculated with R is a table of values containing columns for the predicted values (“Estimate”, abbreviated B), the standard error (“Std. Error”), the t value, and

¹⁴In R, interactions between predictor variables are specified with a colon : or an asterisk *, where $a*b$ is shorthand for $a+b+a:b$.

the p value. The first value in the “Estimate” column is called the intercept. It is the modified mean value for the reference level, which serves as the basis of comparison.¹⁵ In a study on New Englishes with the predictor variables variety and year, this could be the value for the shared parent variety at year 0 (cf. section 4.2.5 on centering numeric predictor variables). This value cannot reach statistical significance, even though the model can indicate a lower-than-chance probability. (In this case, the low p value means that the value is significantly different from 0.) Every subsequent estimate in the table is based on this intercept value and lists the change to the intercept that a predictor variable causes. A positive value means an increase, a negative value a decrease. The absolute value is obtained by adding the estimates pertaining to the relevant predictors.

The standard error is a measure for how sure one can be of the estimate being correct. As Wolk et al. (2013: 401) summarize, “with 95% certainty the true coefficient will lie within the range of the reported coefficient [=estimate] plus or minus twice the SE. If that range does not include zero, the coefficient is statistically significant.” That means that the lower the standard error, the more reliable the estimate is (cf. Bortz 2005: 194). The t value, shown in the third column, is calculated dividing the estimate by the standard error (cf. Baayen 2008: 89–90). On the basis of the t value, the p value is calculated (cf. *ibid.*: 89). The p value indicates how probable the results are under the assumption that the null hypothesis is true. A result is called statistically significant if its occurrence is so improbable that it cannot be attributed to chance. The higher the unlikelihood of occurrence of a result, the more highly statistically significant it is. This is indicated by zero to three asterisks for ‘not significant’ to ‘highly significant’. The significance levels used throughout this study are given in table 4.2 (cf. Gries 2013b: 29).

Table 4.2: Significance levels

significance level	p value	indicated as
highly significant	< .001	***
very significant	< .01	**
significant	< .05	*
marginally significant	< .1	.
not significant	> .1	

¹⁵In the present study, only treatment contrasts are used. Cf. Crawley (2013: 440–442) for a detailed assessment of the benefits and drawbacks of the various contrast types.

4.2.3 Logistic regression

Logistic regressions constitute a subgroup of generalized linear models. Generalized linear models differ from linear models in their method of ascertaining the relation between variables. While linear regression makes use of least squares regression, generalized linear models are calculated using maximum likelihood estimation. With this method, predicted values are calculated until they are “most similar to the observed values” (Baayen 2008: 195). This process is also called “iterative fitting” because the model is calculated over and over again until all predicted values match as closely as possible.¹⁶

Logistic regression is adequate in cases where the dependent variable is not continuous in nature but binary. Examples of such binary variables are true vs. false, success vs. failure etc., that is, variables where only one out of two options can occur. Logistic regression models infer from the input data the degree of variation of a binary variable (i.e. the chances of realization of either one or the other option) depending on several independent (or predictor) variables, i.e. explanatory factors (cf. *ibid.*).

Interpreting the output of a generalized linear model

The output of a generalized linear model is comparable to that of a linear model, with one notable exception. In the “Estimate” column, values are given in logarithmically transformed odds (log odds). The log odds for every variable are given compared to the reference level. The estimate that is indicated for the reference level (intercept) is not a meaningful value. It is neither an absolute value nor can it acquire any statistical significance (even if the model indicates a lower-than-chance probability). Once again, all subsequent estimates in the model are based on this value. Values of estimates can be positive or negative, indicating that a predictor induces either an increase or a decrease in log odds.

Log odds, the form in which estimates are given, are the logarithmic values of odds. Odds are calculated by dividing all success by all failures. Odds O can be calculated from probabilities by dividing the probability for success P by the probability for failure $1 - P$:

$$O = \frac{\text{successes}}{\text{failures}} = \frac{P}{1 - P} \quad (4.6)$$

The probability can be calculated from the odds as shown in the following equation:

$$P = \frac{\text{successes}}{\text{success} + \text{failures}} = \frac{O}{1 + O} \quad (4.7)$$

¹⁶Generalized linear models were fitted with the function `glm()` in R.

Analogous to the linear model, the next column in the output table lists the standard error for the estimated values. Depending on the model, the subsequent column provides the t (linear regression) or z value (logistic regression). The z value indicates how much an estimate deviates from the mean. The higher the absolute z score, the further away from the mean the estimate is. On the basis of the z value it is possible to calculate the p value, again shown in the last column, which indicates in how far a result is likely to be due to chance.

4.2.4 Linear and logistic regression with random effects

Another type of regression that is used in this study is mixed-effects regression (cf. Pinheiro and Bates 2000).¹⁷ Linear mixed models, also called multilevel linear models, are fit by restricted maximum likelihood estimation, while logistic mixed models are fit by maximum likelihood estimation. Wolk et al. (2013: 399–400) summarize the major advantage of this extension of regression models as follows.¹⁸

In addition to so-called ‘fixed effects’ – which are classically estimated predictors suited for assessing the reliability of the effect of repeatable characteristics – mixed-effects modeling allows for ‘random effects’ that are well suited to capture variation dependent on open-ended, potentially hierarchical and unbalanced groups.

In other words, the main advantage of fitting a mixed-effects model is that it can account for so-called random effects, i.e. idiosyncratic effects of elements of a group that should not be considered as contributing to variation. This helps to capture “group-level variation in the uncertainty for individual-level coefficients” (Gelman and Hill 2007: 246). Among the latter figure individual test subjects as in chapter 8, or, as in chapter 6, individual verbs from a group of randomly selected verbs. By including random effects in the regression model, it becomes possible to establish trends that go beyond individual items of a group. Wolk et al.’s (2013) study, for example, focuses on genitive and dative alternation in British and American English. They are interested in general alternation patterns that they try to find independently of specific authors’ preferences. In the studies presented here, random effects are always modeled as random intercepts (cf. Baayen 2008: 85–91, 247).

¹⁷Mixed-effects linear regressions were fitted with the function `lmer()`, mixed-effects logistic regressions were fitted with the function `glmer()`. Both are available in the `lme4` package (cf. Bates et al. 2014).

¹⁸Gries (2015) gives reasons why mixed-effects models are of relevance to particularly corpus-linguistic investigations. Field et al. (2012: 859–860) provide a more general list of the benefits of mixed-effects models.

4.2.5 Frequency as a continuous predictor

In the following statistical analyses, unless otherwise indicated, all numeric predictor variables are normalized, centered and logged. Normalizing values serves to adjust all values to a common scale. This facilitates comparison and is common practice when operating with, for example, corpora of different sizes: 1000 hits in a corpus of 1 million words are considerably more than 1000 hits in a corpus of 1 billion words.

Numeric values are also centered, which means that the mean of the entire set of values is subtracted from each individual value (cf. Gries 2009: 121). That way, values arrange themselves around zero, going both above and below it. Centering is of particular importance in models including interactions between variables, since this process will reduce collinearity. It can additionally be of use if the value 0 of the continuous variable is not meaningful. In the case of word frequency, predictions for a frequency of 0 are rather meaningless, which is why values should be centered in these cases, so that predictions are made for the mean value (cf. Grace-Martin 2014).

Finally, the logarithm of all numeric predictor values representing frequencies is calculated. This is done to account for the skewness of word frequencies. “There are many low-probability words and relatively few high-probability words”, as Baayen (2008: 92) notes. On a linear scale, the distance between 1 and 11 and 1001 and 1011 occurrences of a word in a corpus is 10 in both cases. Nonetheless, corpus-linguistically speaking, the difference between 1001 and 1011 tokens is not as drastic as the one between 1 and 11 occurrences. If a word were to appear just once in an entire corpus, it might even be discarded from the analysis since it could be considered an ad-hoc formation rather than the instantiation of a productive pattern. On the logarithmic scale, these differences are expressed differently in the sense that higher values are ‘closer together’ and lower values are ‘further apart’. As is apparent from table 4.3, the logarithmic scale is more suitable than the linear scale, particularly for corpus-linguistic calculations. The logarithm gives the difference between two values in percentages instead of absolute numbers. An increase in frequency by 10 from 1 to 11 will thus have a higher value ($2.40 - 0 = 2.40$) than an increase by 10 from 1001 to 1011 ($6.92 - 6.91 = 0.01$). Figure 4.2 shows the linear and the logarithmic function.

Table 4.3: Linear vs. logarithmic values

x	1	11	1001	1011
$\ln(x)$	0	2.40	6.91	6.92

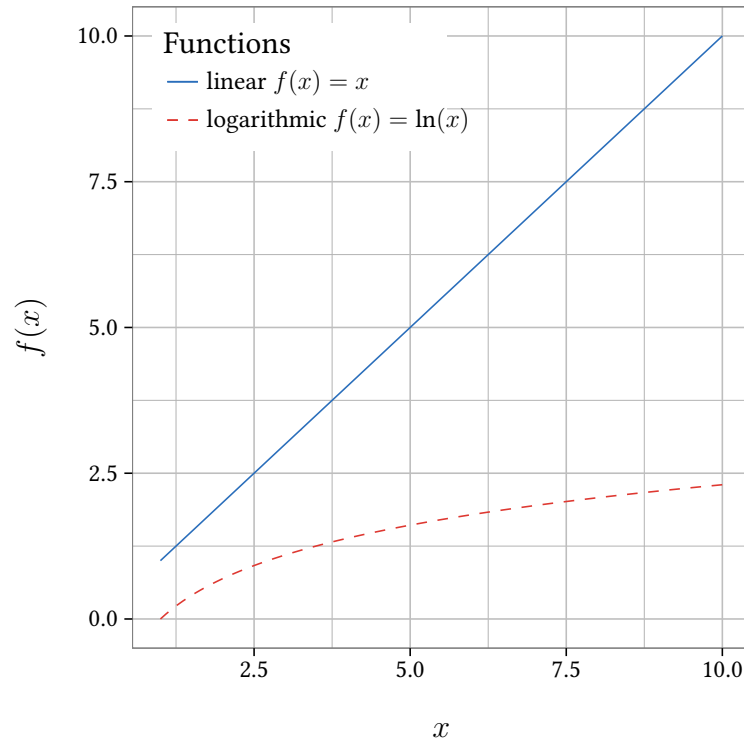


Figure 4.2: The linear and the natural logarithmic function

Logarithmic transformation is also applied to facilitate statistical calculations. As Baayen (2008: 31) points out,

[t]he reason that many of the variables [...] are logarithmically transformed is to eliminate or at least substantially reduce the skewing in their distribution. This reduction is necessary for most of the statistical techniques [...] to work appropriately. Without the logarithmic transformation, just a few extreme outliers might dominate the outcome, partially or even completely obscuring the main trends characterizing the majority of data points.

In the following, frequencies are furthermore always modeled as continuous predictors. That is, frequency is interpreted as a gradual phenomenon, rather than a discrete one that requires classifying items into different frequency bins (e.g. a high- vs. a low-frequency class). This decision is based on Arnon and Snider's (2010: 76) finding that modeling frequency as a continuous variable yields a model of higher explanatory power than modeling frequency as a binary predictor.

4.2.6 Model criticism

Model criticism is an integral part of statistical analysis. A statistical model should always be evaluated critically so as to determine its explanatory power. As Field et al. (2012: 339) humorously point out:

[R]unning a regression without checking how well the model fits the data is like buying a new pair of trousers without trying them on – they might look fine on the hanger but get them home and you find you’re Johnny-tight-pants. The trousers do their job [...] but they have no real-life value [...].

Fitting a regression model

It is often recommended to make use of stepwise regression to arrive at the final regression model that is then reported (cf. e.g. Gries 2013b: 259–261). There are two major methods of stepwise regression, either forward selection or backward selection (cf. e.g. *ibid.*: 260). In forward selection, the final model is gradually built up adding predictors and interactions to the previous model. In backward selection, the maximally complex model is stripped of predictors and interactions between them until arriving at the final model. The final model is reached when adding or discarding predictors does not significantly improve the model any more (cf. *ibid.*). According to Gries (*ibid.*), backward regression “is most widely used in linguistics”. This may be due to the fact that forward regression “runs a higher risk of making a Type II error”, that is of failing to detect significant predictors, as Field et al. (2012: 321) note.

However, there are others who are fairly skeptical of stepwise regression (cf. e.g. *ibid.*: 264–265; Gelman and Hill 2007: 68–69; Harrell 2015: 67–72). Instead, they recommend basing statistical models on theoretical considerations. Harrell (*ibid.*: 95), for example, points out that it is preferable to “[f]ormulate good hypotheses that lead to specification of relevant candidate predictors and possible interactions”. For Gries (2013b: 335), the fact that there is no “widely accepted [...] model selection process” is one of the major drawbacks of mixed-effects models. In the present study, a mixture of both approaches is adopted: A priori theoretical considerations guide the formulation of a maximal model which is then reduced in a second step (backward regression) if the maximal model were to contain predictors that turn out to be non-significant. From one step to the next, the ‘old’ and the ‘new’ model are contrasted analyzing various information criteria.

Choosing a model by means of information criteria

There are a number of quality measures that can be drawn on to compare regression models. This study relies on the AIC and the BIC. The AIC, the *Akaike Information Criterion* (cf. Akaike 1973, 1974), is a “parsimony-adjusted measure of fit” (Field et al. 2012: 263), that is, it gives a coefficient that calculates the model fit penalizing the addition of more predictors. This means that adding a predictor will only yield a considerably better AIC if this predictor significantly benefits the fit of the model. Larger values indicate worse fit, smaller values indicate better fit. The AIC is a relational measure, i.e. only models fitted to the same data can be compared by means of the AIC (cf. *ibid.*). The BIC, *Schwarz’s Bayesian Criterion* (Schwarz 1978), “is the same as the AIC but adjusts the penalty included in the AIC [...] by the number of cases” (Field et al. 2012: 263). The smaller the BIC is, the better the model fit is.

Evaluating a model by means of comparison to the baseline model

A common procedure to test whether calculating a logistic regression model is justified is by evaluating how the number of correct predictions changes with the calculation of the model. The statistical model is usually tested against a baseline model that will always predict that the most likely thing is going to happen. If the regression model leads to an increase in prediction accuracy compared to the baseline model, it is better than the baseline model (cf. Gelman and Hill 2007: 99–100). Figure 4.3 illustrates in a simplified way how conversion plays out in reality (a), according to the baseline model (b), and as predicted by the logistic regression model (c). The shaded areas indicate those cases where the model over- or underestimates the odds of either possible outcome.

The study in chapter 6 serves as an illustration. The logistic regression model in this specific case calculates the chance that a deverbal noun is realized as a converted form or as a derived form. In the following, n is to represent tokens identified as nouns (e.g. *imagine*) and a is to stand for tokens of the deverbal nominal alternative (e.g. *imagination*). The probability of conversion P_{real} in the data is then calculated as follows:

$$P_{\text{real}} = \frac{\sum n}{\sum (n + a)} = 0.00162 \quad (4.8)$$

The baseline model will assume that the probability of conversion is 0 and the probability of the deverbal noun is 1, because the latter is the more frequent option and the baseline model always predicts the most frequent option. The baseline model will hence be correct

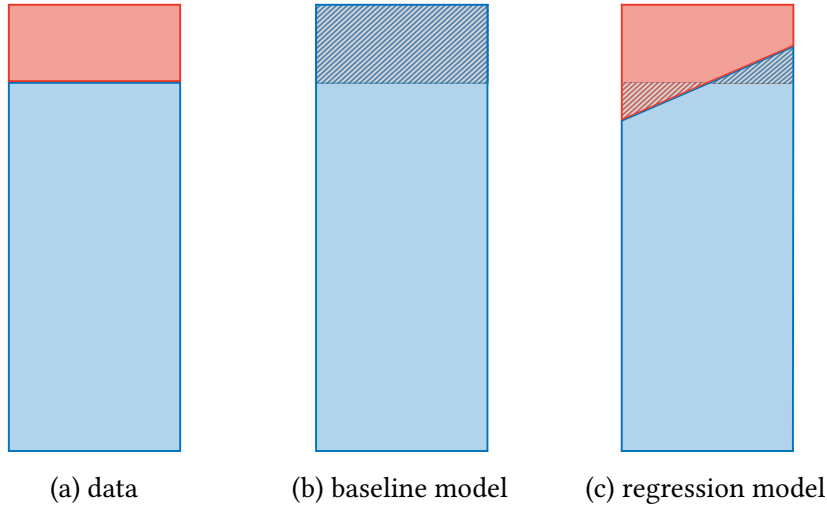


Figure 4.3: Proportion of possible outcome A (blue) to possible outcome B (red) in the data (a), as predicted by the baseline model (b), and by the regression model (c). Shaded areas indicate prediction errors.

in $(1 - P_{\text{real}}) \cdot 100\% \approx 99.8\%$ of all cases.¹⁹ Its prediction error e_0 , i.e. the probability of predicting an error, is thus equal to the probability of conversion in the data.

$$e_0 = P_{\text{real}} = 0.00162 \quad (4.9)$$

If the prediction error is smaller for the logistic regression model, i.e. if the model can improve the prediction error by predicting more accurately, then computing the model makes sense. The prediction error of the model is calculated by summing up the deviance of the prediction \hat{n} from the actual count for each individual nominal token n and then dividing this sum by the sum of all tokens:

$$e_1 = \frac{\sum |n - \hat{n}|}{\sum (n + a)} = 0.00082 \quad (4.10)$$

Since $e_1 < e_0$, the logistic regression model in chapter 6 predicts the real distribution of conversion better than the baseline model.

¹⁹This shows that verb-to-noun conversion, where the converted form is in competition with an established derived noun, is indeed a very rare phenomenon that requires the use of large corpora in order to detect it, let alone establish gradual differences between varieties.

4.3 Experimental methods

In order to corroborate the findings obtained through the analysis of corpora, an experiment is conducted (cf. chapter 8). The integration of corpus-linguistic and experimental methods has previously proven to be a highly effective combination in various sub-disciplines of linguistics²⁰ and is therefore also applied in the present case. (A more theoretical discussion of the benefits of combining corpus analytic and experimental methods is provided in e.g. Gilquin and Gries 2009: 9 or Schönefeld 2011: 22–24.)

4.3.1 Acceptability judgment

A common experimental method in linguistics is grammaticality or acceptability judgment.²¹ Schütze and Sprouse (2013: 28) define acceptability judgments as follows:

Acceptability judgments [...] involve explicitly asking speakers to ‘judge’ (i.e., report their spontaneous reaction concerning) whether a particular string of words is a possible utterance of their language, with an intended interpretation either implied or explicitly stated.

The roots of linguistic judgments lie in the Chomskyan competence/performance dichotomy. Through explicit grammaticality judgments, it was thought possible to access the competence of native speakers, while acceptability judgments were considered reflections of performance (cf. Schütze 1996: 20). (For an overview of and a critical comment on the history of grammaticality and grammaticality judgments the reader is referred to *ibid.*: 19–53.) The acceptability judgment proposed by Schütze (*ibid.*) differs from the judgment of grammaticality in the Chomskyan sense in that it aims at the “spontaneous reaction” of the speaker and discourages the explicit accessing of metalinguistic knowledge (Schütze and Sprouse 2013: 28).

Acceptability judgments rely on the perceived acceptability of stimuli (cf. *ibid.*), that is, the measure for acceptability is highly subjective. Nonetheless, acceptability judgments are a valuable tool in the linguist’s experimental tool kit. The main reason for judgment data is that “failure to appear in even a very large corpus (such as the Web) is not evidence for ungrammaticality, nor is appearance evidence for grammaticality”, as Schütze and Sprouse

²⁰See Boyd and Goldberg (2011), Bresnan (2007), Goldberg (2011), Gries et al. (2005), and Suttle and Goldberg (2011) for studies from Cognitive Construction Grammar, Brandt and Kidd (2011), Meunier and Littré (2013), Siyanova and Schmitt (2008), and Wulff (2009) for language acquisition studies, or Gries (2002) and Lorenz (2013) for studies on language variation and change.

²¹It has been argued (cf. e.g. Chomsky 1965: 10–11) that grammaticality and acceptability are not the same. Schütze (1996: 19–53) offers a detailed account of both concepts. Here, the notion of acceptability judgment is preferred.

(ibid.: 29) argue. Thus, it makes sense to complement a web-corpus study with an acceptability judgment task so as to find out whether the results of both studies tally.

The advantages of this type of task are, first, that it allows to obtain responses to infrequent language phenomena that do not occur often in natural conversation (cf. Schütze 1996: 29). Second, it is a simple procedure that is easily implemented. Third, its results “are highly systematic across speakers” (Gibson et al. 2011: 512). Furthermore, in the experimental setting it is possible to disentangle features characteristic of spoken speech (“slips, unfinished utterances”) from ‘pure’, “grammatical” language production (Schütze 1996: 2). Additionally, because the experimental stimulus is devoid of a communicative purpose, it is possible to obtain reactions to the stimulus without the communicative context interfering (cf. ibid.).²²

The main arguments against acceptability judgment are that linguistic intuition is not to be trusted (Householder 1965 in ibid.: 3–4), that the judgment behavior is artificial and has little in common with natural linguistic behavior, and that the stimuli are often artificial (cf. Bresnan 2007: 91). All these points are addressed in the acceptability judgment experiment in the present study. In order to reduce the chance that speakers access their metalinguistic knowledge, the instructions do not explicitly point the participants in the direction of grammar or ‘correctness’. The scale on which stimuli are to be rated is consequently not dichotomous (correct/incorrect) but continuous, so as to account for the gradient nature of grammaticality and acceptability. Second, the stimuli are not artificial but have been taken from the corpus, as recommended by Bresnan (ibid.). While one could rightly argue that the occurrence of a sentence in a web corpus does not mean that it was produced by a native speaker of a specific variety (cf. section 4.1.3 on the challenges in the compilation of GloWbE), it is assumed that the majority of sentences indeed is. Both measures, the hedged formulation of the instructions and the naturalness of the stimuli, are hypothesized to suggest a linguistically fairly natural environment.

Likert scale task

Schütze and Sprouse (2013: 31–36) describe various types of acceptability judgment tasks. The one adopted in the present study is what is called the “Likert scale task” (ibid.: 33). The participants’ task is to “locate a sentence on a numerical scale”. This task is fairly intuitive in that the higher the number, the more acceptable the stimulus is deemed to be. Due to the

²²Generally, it can be more difficult to make meaning of an utterance if the context is not given. A linguistic phenomenon that might be judged acceptable in a specific context could easily be judged unacceptable without the context (e.g. in the case of the constructions such as *?Ed hammered the metal safe* that Boas 2011 describes). By presenting the phenomenon without context, these context effects can be filtered out.

drawbacks associated with a conventional Likert scale (cf. Schütze and Sprouse 2013: 33),²³ in this experiment a continuous scale ranging from 0 to 1000 is adopted (cf. section 8.1).

As far as task design is considered, Schütze and Sprouse (*ibid.*: 39) recommend the use of filler items. The two main reasons for the use of fillers are that, first, with fillers interspersed with the actual test items participants are less likely to become aware of the purpose of the experiment. Second, fillers in judgment tasks can be designed in such a way that they require the use of the entire response scale. In doing so, it is possible to minimize a potential “scale bias, which occurs when one participant decides to use the response scale differently from other participants, such as only using one end of the scale (skew), or only using a limited range of responses (compression)” (*ibid.*).

4.3.2 Reaction time

The measurement of reaction times (RT, also response time or response latency, cf. Jiang 2012: 2) is a common method in psychology (cf. Hergenhahn and Henley 2014: 254–255). Reaction times are generally taken to reflect processing speed, with smaller reaction times indicating faster processing, most likely resulting from less processing effort. As Jiang (2012: 2) explains,

[t]he use of RT data is based on the premise that cognitive processes take time and by observing how long it takes individuals to respond to different stimuli or perform a task [...], we can [...] infer about the cognitive processes or mechanisms involved in language processing.

The main advantage of RT experimentation is that it is applicable to any cognitive process because “[a]ny mental event takes time” and thus lends itself to a measurement of RT (*ibid.*: 7). Furthermore, the measurement of RT is a comparatively simple way of gaining insight into on-line processing (cf. *ibid.*: 10). In contrast to other ways of experimentation such as grammaticality judgment, RT experimentation “can often minimize the involvement of explicit knowledge” (*ibid.*: 9), with “explicit knowledge” including (meta-)linguistic knowledge.

Despite these advantages, RT experimentation has two major inconveniences. The first is that “RT is not linguistic behavior itself” (*ibid.*: 12). That is, any conclusion that is drawn from the results of RT experimentation and any theoretic linguistic insight which might subsequently be gained is based on inferences that the researcher has made on the grounds of

²³One crucial issue is the discreteness of the scale. Sentences perceived as deserving a rating of 1.5 and 2.5 respectively might both end up being rated as ‘2’, whereas in actual fact the participant does perceive a distance of 1 between the sentences.

mere differences in how fast participants responded to a stimulus in an experiment. Secondly, RT is a measure that is highly susceptible to many influences. The factors affecting reaction time not only include the type of experiment, the type of stimulus or the stimulus intensity (Kosinski 2013), but are extremely diverse, as the list below shows. All these factors (and various other, less relevant factors) are described in further detail in Kosinski (ibid.).

stable internal factors age, gender, preferred hand, personality type, intelligence

variable internal factors arousal, fatigue, illness, alcohol, stimulant or depressant drugs

external, task-related factors practice, previous errors, order of presentation of items, distraction

Due to the fact that controlling for all these variables is extremely demanding, in the present study, RT is only taken as a relational measure. In other words, no conclusions will be drawn from the absolute length of reaction times, but only from an increase or decrease in RT to a certain stimulus compared to a baseline RT. Faster RTs can be seen as indicators of faster processing that results from a greater familiarity with the stimulus. A less familiar stimulus will lead to prolonged RTs.

4.3.3 Web-based experimentation

As far as the environment in which experiments on varieties of English are to be conducted is concerned, the world-wide web is of particular interest. The web provides researchers with comparatively easy access to a large number of potential participants at a very low cost, compared to a traditional lab setting. These and various other advantages but also drawbacks as well as the question of the quality of experimental data obtained on the web are laid out in the following.

Potential and limitations of web experiments

The advantages of conducting an experiment on the internet are readily at hand.²⁴ They can be grouped into resource-related, method-related, sample-related advantages. Regarding the resources, web-based experimentation is a method that is fast and requires fewer financial resources than experimentation in a laboratory setting, as it saves the experimenter the cost of running the lab. In addition, an experiment on the internet is always available, which is an advantage in a study of varieties, considering that participants will most likely live in

²⁴For an exhaustive list of advantages cf. Reips (2002: 245), only the most relevant points are discussed here.

different time zones and thus will want to access the experiment at different times of day (cf. Birnbaum 2004b: 361–362).

Furthermore, as regards the method, the distance between the researcher/experimenter and the test subjects generates a certain anonymity which makes it almost impossible for the experimenter to introduce a bias (cf. Birnbaum 2004a: 822; Schnell et al. 2011: 369–377; Schnoebelen and Kuperman 2010: 463).

As far as the sample is concerned, the internet allows the researcher to recruit (with comparative ease) a large number of participants (cf. Birnbaum 2004b: 361–362). The latter will usually form a more diverse population than the usual undergraduate student population upon whom experimenters often rely (cf. Birnbaum 2004a: 813; Reips 2002: 245). Both the size of the sample and its diversity lead to a higher statistical power of the results and to a better generalizability (cf. *ibid.*). Moreover, web experiments provide comparatively easy access for the participants (cf. *ibid.*; Schnoebelen and Kuperman 2010: 463), who generally show a high degree of voluntary participation (cf. Reips 2002: 245).

Nonetheless, web-based experimentation also comes at a cost. The main drawback is the lack of control over both the sample and the process of conducting the experiment. On the web, participants cannot be sampled as accurately as in a laboratory setting. Schnell et al. (2011: 371) point out that the “convenience sample” which the web provides is not representative, and therefore does not allow for generalizations. However, considering that the GloWbE corpus, the main corpus tool for the present study, is compiled of texts produced by internet users, it can be assumed that there is a certain degree of overlap in characteristics between those speakers that have produced the corpus texts and those that take part in the experiment. Another limitation is that the sample might show a self-selection bias, meaning that only those who really want to participate do in fact participate in the experiment (cf. Reips 2002: 245; Schnoebelen and Kuperman 2010: 462). Yet, it has to be acknowledged that a self-selection bias might also apply to the lab sample (cf. *ibid.*).

The other domain over which the experimenter lacks control in web-based experimentation is the question of how the individual participant takes the experiment. This involves the circumstances (e.g. distractions, presence of other people, use of ‘forbidden’ help such as a dictionary or advice from another person, cf. Birnbaum 2004b: 361–362) but also each participant’s attitude towards the experiment. The anonymity of the web may tempt participants not to comply with the instructions of the experiment and answer questions with a low accuracy to ‘get it over with’ (cf. Eickhoff and de Vries 2013: 121), or to drop out before they have completed the experiment (cf. Birnbaum 2004b: 375; Reips 2002: 245). While this does not constitute a violation of the instructions of the experiment, it can seriously compro-

mise the results, as Birnbaum (2004a: 817) demonstrates. A further drawback is that there is no method to completely exclude the possibility of multiple submissions (cf. *ibid.*: 813–816; Reips 2002: 245). Nevertheless, Birnbaum (2004a: 813–816) notes that multiple submissions do not occur frequently. Finally, due to the remoteness of participant and experimenter, the former cannot ask clarification questions (cf. *ibid.*: 822). Considering that the correct interpretation of the instructions is vital for the quality of the data, Crump et al. (2013: 17) recommend to check whether participants have understood the instructions before proceeding with the experiment.

Since cheating and a high drop-out rate compromise the experiment most seriously, these points are addressed more explicitly in the following.

Deceptive and inattentive behavior

The anonymity of the web may tempt some participants to seek distractions while they are participating in an experiment or to optimize their working routine through completing tasks as quickly as they can without paying careful attention to the accuracy of their answers. This tendency could increase in situations where participants can win a prize or receive payment.

However, there are ways of detecting deceptive behavior. For closed-class questions such behavior can easily be filtered out by comparing the answers to a gold standard (cf. Eickhoff and de Vries 2011: 12). For open-class questions it is helpful to check whether a participant has repeatedly entered a “generic string of words” (cf. *ibid.*). Mason and Suri (2012: 14) further propose to reduce cheating by introducing CAPTCHA questions.²⁵ Another way of filtering out “low-quality responses” is to analyze the answering patterns, as “low-quality responses usually show low-entropy patterns” (*ibid.*). Generally, it is advisable to design the task in such a way that the task in itself discourages deceptive behavior. Eickhoff and de Vries (2011: 13) found that higher task complexity reduced cheating by approximately 17%.

It has also been suggested that researchers make use of so-called Instructional Manipulation Checks (IMC) or screeners (cf. Berinsky et al. 2014; Oppenheimer et al. 2009). According to Oppenheimer et al. (2009: 867),

[an IMC] consists of a question embedded within the experimental materials that is similar to the other questions in length and response format (e.g. Likert scale, check boxes, etc.). However, unlike the other questions, the IMC asks participants to ignore the standard response format and instead provide a confirmation that they have read the instruction.

²⁵CAPTCHA is an acronym for *Completely Automated Public Turing test to tell Computers and Humans Apart*. CAPTCHA questions can easily be answered by humans but not by computers, e.g. by copying letters or numbers presented as an image into a blank.

The main assumptions that underlie the use of IMC or screener questions are that “(1) participants who fail the IMC also fail to follow other instructions in the survey; and (2) failing to follow these other instructions will result in less reliable and valid data” (Oppenheimer et al. 2009: 868). In Oppenheimer et al.’s (ibid.: 869) study, 46% of all participants failed the IMC, which indicates that the noise introduced into the data by inattentive respondents can be quite large. This clearly shows that there is a need for determining whether participants are paying attention or not so as to avoid skewed results. Yet, the appropriate number of IMCs as well as the impact of IMCs on participants are still disputed (cf. Berinsky et al. 2014; Oppenheimer et al. 2009). What is more, while Oppenheimer et al. (2009: 871) found no demographic differences between those participants who passed and those who failed the IMC, Berinsky et al. (2014: 749) report age, gender and race to influence the failure of an IMC. Consequently, excluding data by participants who fail IMCs could entail introducing a considerable population bias.

Even though IMCs may undoubtedly be helpful for certain tasks, in the present experiment they are not used. The reasons are twofold. The first reason is that one of the tasks, the maze task (cf. section 8.2), does not require a screener. The main aim of the experimenter should be to design the experiment in such a way that it discourages inattentive behavior in the first place (cf. ibid.: 752). This is the case for the maze task. Due to the complexity of the task, it is impossible for the participants not to be attentive. For the other task, an acceptability rating task (cf. section 8.1), it would make sense to ensure that participants are reading the sentences that are to be rated with careful attention. Nonetheless, explicitly making participants aware of the fact that their level of attention is monitored might induce in them a tendency to comply with language norms so as to prove that they are, in fact, paying attention. Answering truthfully and stating that one finds the sentence acceptable/unacceptable even if it contains non-standard features might be interpreted as being an indicator of inattentive behavior. In order to avoid such (probably even unconscious) reasoning on the part of the participant, screener questions are not employed in the rating task.

In summary, in the present context the implementation of an IMC does not seem to be a worthwhile undertaking. While it might help filter out inattentive participants, it might also discourage others from answering truthfully, which would also result in data skewing. It is therefore necessary to adopt covert strategies such as the measuring of response times—assessing whether there are participants who show implausibly short or long overall completion times—or to check for implausible deviations in response behavior, e.g. an inconsistent rating of clearly standard/non-standard sentences.

Drop-out

Another challenge in web-based experimentation is the large drop-out rate compared to lab settings. In Dandurand et al.'s (2008: 431) web-based study, out of 600 people who clicked on the link to the study and saw the first page of it, only 27 (4.5%) completed it. While this might be an extreme case, it aptly shows that it is necessary to develop strategies to motivate participants to begin and then finish the experiment once they have started it.

One possibility is to introduce a reward in the form of a payment or a prize. Frick et al. (2001: 214) could show that introducing a lottery condition highly significantly reduced the drop-out rate from 18.5% to 9.5% but did not affect the quality of the responses to their survey questions. However, they further found that “the lottery information does not result in additional motivation to start with the experiment, but diminishes drop out tendency caused by other factors” (ibid.: 217). In Musch and Reips's (2000: 80) study, the difference in completion rate depending on payment was even more striking, with a completion rate of 86% with a reward and 55% without a reward.

Another option is to adopt the *high-hurdle technique* (cf. Reips 2002: 249). The idea behind this technique is to put “motivationally adverse factors” at the beginning of the experiment, for example, by asking participants for personal information early on, making them read long texts first and then diminishing the amount, etc. This will prompt the unmotivated participants to quit the experiment at an early stage, causing less drop-out over the course of the experiment. Frick et al. (2001: 215), for instance, found that asking for personal information “early in the experiment” significantly reduced the drop-out rate (from 17.5% to 10.3%).

Reips (2002: 249) further proposes to include a warm-up phase before the actual experiment starts. The warm-up phase could include practice trials or attention checks, or familiarize participants with behavioral routines. Postponing the start of the experiment in this way will once again lead to unmotivated participants dropping out early.

Data quality

The most important point to assess before conducting an experiment over the internet is the question of how the data gathered on the web compare to data from the lab, and whether the web is a valid resource for research. Many studies have demonstrated that web data and lab data do indeed exhibit a very high level of agreement (cf. Birnbaum 2004a: 824–827; Krantz and Dalal 2000: 56; Musch and Reips 2000: 81–82). McGraw et al. (2000) could show that the results of several psychological experiments conducted on the web (word recogni-

tion, mental rotation experiment, Stroop test, priming test) replicated the results obtained in the traditional lab setting. A study by Dandurand et al. (2008: 432) suggests that the time required for a complex problem-solving task was not different on the web than in the lab when corrected for participants' age.

A recent trend in web-based research is to recruit participants through crowdsourcing platforms (see below). There are a number of studies that indicate that the crowdsourcing environment also provides highly reliable data that are very similar to the results from laboratory studies. Crump et al. (2013: 17) found that various reaction time measurements (Stroop effect, task-switching cost, Flanker effect, Simon effect) could be replicated on crowdsourcing platforms and that the data “compare[d] well to laboratory studies”. For the rating of speech data McAllister Byun et al. (2015: 78) observed that there was “strong agreement [...] between AMT listeners [viz. listeners recruited via Amazon Mechanical Turk, a large crowdsourcing platform] and experienced raters”. Kuperman et al. (2012: 987) reported high correlations for age-of-acquisition ratings. While crowdsourcing differs from ‘regular’ web-based experimentation in that participants receive remuneration for their work, it can be assumed that once motivated participants without an intention to deceive have been recruited for a web-based experiment, their results compare to data gathered in the lab. According to McGraw et al. (2000: 504), the major advantage in web-based experimentation is the large size of the sample; as they say, “numbers will swamp noise”.

Crowdsourcing

In order to provide test subjects with an extrinsic motivation to participate in the experiment, the experiment in this study was designed to be set up on a crowdsourcing platform. The term *crowdsourcing*, in analogy to *outsourcing*, was coined in 2006 (cf. Howe 2006) to describe the act of delegating (usually small and simple) tasks to a large, anonymous crowd of workers who receive small amounts of money in exchange for their work.

The potential of crowdsourcing platforms such as Amazon Mechanical Turk or Microworkers for academic research purposes has already been explored in various fields. In variationist linguistics, however, it is a method that has (not yet) received a lot of attention. A first approach is described by Zaidan and Callison-Burch (2014), who asked workers on Amazon Mechanical Turk to identify texts as being written in dialectal or standard Arabic. Using a crowdsourcing platform for an experiment to explore varieties of English is thus not only both financially attractive and time-efficient. It would also mean covering new methodological ground.

4.4 Summary: Integrating quantitative and qualitative analyses

Nonetheless, crowdsourcing turned out to be unfeasible in the present context. Amazon Mechanical Turk, the platform that would have been chosen for having received most attention in the academic community, does not have enough workers from Hong Kong and Singapore registered. The same applies to Microworkers, another platform that has previously been used for academic purposes (cf. e.g. Hoßfeld et al. 2013).

Instead, the *friend-of-a-friend* approach was adopted (also called *snowball technique*, cf. Milroy 1980: 46–56; Milroy and Gordon 2003: 32). “This approach utilizes the social networks of participants in the study to recruit potential new participants” (ibid.). Instead of approaching potential participants directly, contact is established via a mutual friend of the researcher and the potential participant. That way, the researcher is not perceived as a complete stranger but as a friend of a friend, which might lead potential participants to show behavior that is characteristic of friendship. More precisely, potential participants may “feel some obligation to help” (ibid.: 75) and might therefore sign up more readily. A further advantage of this approach is that it encourages the use of everyday language, whereas approaching participants “through individuals with a clear institutional status [...] can often lead to rather standardized speakers [and speech]” (ibid.). It can be assumed that the *friend-of-a-friend* approach is practicable in all the regions investigated. Apart from Milroy’s (1980) there are numerous other studies that have demonstrated that the scheme works well in westernized countries (cf. e.g. Tagliamonte 2006: 22). As regards non-western politeness norms, Siemund et al. (2014: 348) observed that the approach is even more effective in Singapore.

4.4 Summary: Integrating quantitative and qualitative analyses

Quantitative methods such as statistical analyses offer many advantages. They are fairly objective and not very susceptible to the researcher’s biases.²⁶ Furthermore, they allow to make generalizations based on large data sets to infer whether language phenomena are of statistical significance, i.e. whether their probability is larger than simply due to chance, which is probably the only way to make sense of large corpora such as COCA or GloWbE. Quantitative methods hence counteract the disadvantages of qualitative analyses. Qualitative analyses are generally more prone to the researcher’s biases and expectations and they do not allow the researcher to generalize and recognize trends to the same extent.

²⁶Notwithstanding, there are ways for the researcher to deliberately or unconsciously modify statistical analyses so that the results are greatly distorted. Simmons et al. (2011: 1360) describe how their test subjects’ age ‘depended’ on what song they had heard right before indicating their date of birth. They trace such distortion of results back to the researcher’s “flexibility in data collection and analysis” (ibid.: 1359). While this is a rather extreme example, it has to be noted that even the seemingly most objective method can be influenced by the researcher’s bias.

On the other hand, qualitative analyses provide opportunities for detailed, in-depth investigation that quantitative analyses cannot provide. In using quantitative methods, interesting or surprising findings might be overlooked simply because they do not occur with the frequency necessary to affect the statistical model. Furthermore, the input for statistical analyses in most cases has to fit certain pre-defined categories of analysis, which can prove too coarse to reflect actual language use. In the analysis of parts of speech, as is to be done here, categorizing forms into the verb or noun category might prove particularly complex, especially in ambiguous contexts, since it is well known that word classes are gradient with more prototypical and less prototypical elements (cf. e.g. Bauer 2003: 95–96; Crystal 2004; Quirk et al. 1985: 90).

Quantitative analyses can consequently be said to provide the bigger picture, whereas qualitative analyses help take a closer look and investigate specific instances of language use. It is therefore indispensable to combine both approaches. As Gries (2009: 4) states: “quantitative and qualitative methods go hand in hand: qualitative considerations precede and follow the results of quantitative methods [...]. Often a quantitative study allows to identify what merits a qualitative discussion.” This combination of quantitative and qualitative data analysis is adopted in the following chapters dealing with verb-to-noun conversion in native and new varieties of English.

5 Conversion as a productive process in US English

In order to analyze verb-to-noun conversion in new varieties of English, it is a sensible undertaking to first scrutinize verb-to-noun conversion in native varieties of English. This chapter therefore describes verb-to-noun conversion in US English (USE). The choice of USE as a foil might be startling from a historical perspective, due to the fact that BrE is the parent variety of almost all Asian varieties. However, the reasons for choosing USE are conceptually as well as methodologically well-grounded. Firstly, the US variety can be considered the most influential of the native varieties (cf. Mair 2013a: 261, cf. section 1.1.3). It is the “hub of the World System of Englishes” (ibid.), which is why it seems likely that the same process in other varieties is influenced by this variety. Secondly, from a methodological point, the *Corpus of Contemporary American English* (COCA) is the largest monitor corpus of an established variety that is available at this point. Since the aim of this study is to understand the emergence and diffusion of converted forms, a monitor corpus provides valuable insight that other types of corpora comprising texts from only one point in time (such as the *British National Corpus*) cannot provide. In the following, the process of conversion from verb to noun is illustrated drawing on selected English verbs. The major part of the chapter is concerned with the development of the verb/noun DISCONNECT. Subsequently, other exemplary verbs are analyzed in order to outline different paths that conversion can take.

5.1 DISCONNECT VS. CONNECT

The first case in point is the emergence of the noun DISCONNECT in USE. It has been chosen for this case study as DISCONNECT is fully established as a noun. It was not commonly used as a noun until relatively recently but, as will be shown, has passed through the full cycle of the conversion process within a very short period. Formally, it has a singular and a plural form. Additionally, it is fully syntactically functional and not restricted to light verb constructions such as [*have a N*] or [*take a N*]. Furthermore, it is a relatively recent example of

conversion which originated in USE (cf. *OED Online* n.d., July 2015). Moreover, the attested¹ meanings of DISCONNECT (N) can be delimited very clearly, which is ideal for a semantic analysis. The first meaning is the literal meaning that derives from the verb: “1. An act or instance of disconnecting something; *esp.* a break of an electrical or telephone connection.” This meaning was first attested in 1905. The second meaning, first attested only in 1982, is metaphorical and derives from the first meaning: “2. A lack of consistency, understanding, or agreement; a discrepancy.” The subsequent sections are dedicated to exploring the spread of DISCONNECT as a noun as well as its syntactic and semantic functions.

The spread of DISCONNECT (N) is hypothesized to be facilitated by the usage frequencies of both the verbal form and the synonymous derivation. In line with usage-based theory, a comparatively low frequency of occurrence of DISCONNECT as a verb will lead to the string “disconnect” not being associated strongly with a particular part-of-speech (cf. Ungerer 2002: 560–563). Syntactic re-categorization, i.e. a change in word class, is likely to happen. Secondly, the synonymous, deverbal noun DISCONNECTION also influences the rise of DISCONNECT (N). Frequency directly contributes to the success or failure of pre-emption. A relatively low frequency of DISCONNECTION compared to DISCONNECT (N) will not preempt DISCONNECT (N) from gaining more ground.

5.1.1 The ‘rise’ of DISCONNECT (N) and the blocking of CONNECT (N)

The ‘rise’ of DISCONNECT is illustrated with examples from COCA. CONNECT, the antonym, is chosen as a basis of comparison in order to illustrate opposite tendencies. Choosing this verb pair has two main advantages. On the formal side, CONNECT and DISCONNECT show the same form except for the prefix {DIS-}. In both cases, the nominalization through derivation is accomplished by the suffix {-TION}. Etymological similarity (Latin origin) helps exclude effects that might arise from formal constraints on conversion. Furthermore, on the semantic side, antonyms can be expected to be used in similar semantic fields. This excludes potential effects that the semantic field might have on the success or failure of conversion.²

¹For the purposes of this study, a meaning is considered attested when it has entered the *Oxford English Dictionary* (OED, online edition). The rationale for drawing on the OED is explained in footnote 1 on page 138.

²Yet, choosing this pair cannot help disentangle the effects that prefixation, the number of syllables or the stress pattern might have. Therefore, the analysis of (DIS-)CONNECT is complemented by further analyses of other verbs, which are presented in section 5.2. Nonetheless, as the examples in this chapter are merely case studies, they cannot replace a comprehensive investigation of constraints on conversion.

Data³

For the analysis, all forms of the lemmata *CONNECT* and *DISCONNECT* were looked up in COCA, i.e. infinitive or noun in singular (*(dis-)connect*), third person singular form or noun in plural (*(dis-)connects*), present and past participle of the verb (*(dis-)connecting*, *(dis-)connected*). Furthermore, the competing, assumedly synonymous words *connection* and *disconnection* (including plural forms) were searched for.⁴ All tokens of *(dis-)connect* and *(dis-)connects* were then coded for part-of-speech (verb or noun) both mechanically and manually. Some tokens could not be classified as either because the context did not allow for a classification as verb or noun or because the form in question was part of a proper noun such as *Facebook Connect* or *Adobe Connect*. Tokens in which the form in question acted as modifier in a compound (e.g. *disconnect signal*, *connect rates*) were excluded as well. Frequencies of occurrence of the verb (including participles), the noun and the alternative derivation were calculated for each year.⁵

Linear model

The resulting frequencies were input into a linear regression so as to determine whether the increase in the use of the noun *DISCONNECT* compared to the other nominal form and also compared to the verb is significantly higher. The dependent variable in this regression model is the frequency and the predictors for the frequency are the lexeme (converted noun, verb, alternative deverbal noun) as well as the year. The model equation is reproduced in 5.1.⁶

$$\text{frequency} \sim \text{lexeme} * \text{year} \quad (5.1)$$

Table 5.1 shows the results of the linear model. The values in the estimate column represent the estimated change in frequency per million words. Figure 5.1 displays the frequencies for *DISCONNECT* as a noun and a verb and for *DISCONNECTION*. Frequencies per million words for every year from 1990 to 2011 are represented by dots. Over the years, there is a highly significant rise in the use of the noun *DISCONNECT*, whereas the frequencies of the verb and especially of the nominal alternative, *DISCONNECTION*, do not show such a steep increase. In 1990, *DISCONNECT* is mainly used as a verb, the frequency of use of the verbal form is almost

³This procedure was adopted for all case studies in this chapter.

⁴Considering that the case of *(DIS-)CONNECT* and *(DIS-)CONNECTION* is highly unlikely to constitute an instance of total synonymy, the terminology of ‘alternative (deverbal) noun’ is adopted to refer to the suffixed form.

⁵The year 2012 was excluded from the analysis since the corpus is considerably smaller for that year than for the other years. It might well be the case that this section of the corpus is not as accurately balanced as the other sections, which could distort the results.

⁶The command in R is: `lm(frequency ~ lexeme * I(year-1990), data = verb)`. Since this model is a very simple model, stepwise regression procedures as described in chapter 4 were not deemed necessary.

Table 5.1: Linear model for DISCONNECT

	Estimate	Std. Error	z value	p	
(Intercept)	-0.22	0.15	-1.45	0.151	
verb	3.05	0.21	14.34	0.000	***
deverbal noun	0.59	0.21	2.78	0.007	**
year	0.18	0.01	14.49	0.000	***
verb : year	-0.09	0.02	-5.20	0.000	***
deverbal noun : year	-0.16	0.02	-9.06	0.000	***

three times the frequency of nominal uses. After 22 years, the frequency of nominal DISCONNECT has augmented spectacularly, leaving behind the nominal alternative DISCONNECTION.

This is reflected in the linear model (cf. table 5.1).⁷ The increase of nominal DISCONNECT over time is highly significant ($p < .001$). For every year, the frequency per million words increases by 0.18. The developments of the verbal form and the alternative form differ significantly from that of the nominal form. Their frequency of use increases to a smaller extent, as is visible from the corresponding estimates ($0.18 + (-0.09) = 0.09$ for the verb and $0.18 + (-0.16) = 0.02$ for the nominal alternative). In recent years, the frequency of use as a verb and as a noun has converged, with the verbal form occurring not even twice as often. DISCONNECTION shows a subtle increase, most likely due to the increase in frequency of the entire word family. These developmental trends are also visible in the lines in figure 5.1, which correspond to the regression lines. The line for DISCONNECT (N) shows the steepest slope, which indicates that out of the three forms, DISCONNECT as a noun has experienced the greatest increase.

Generalized linear model

A generalized linear model was subsequently calculated so as to establish the influence of various independent predictors on the odds of conversion from verb to noun. While the linear model above calculates the development of frequency over time of every form independently of other forms, the generalized linear model serves to reveal dependencies between the forms,

⁷The first three lines of table 5.1 cannot be interpreted in a meaningful way. The intercept in itself has no meaning. The other two lines suggest that there are considerably more tokens for the verb and a slightly higher number of tokens for the alternative than for nominal DISCONNECT. This information in itself cannot contribute to answering the question of how these forms have developed between the years 1990 and 2011.

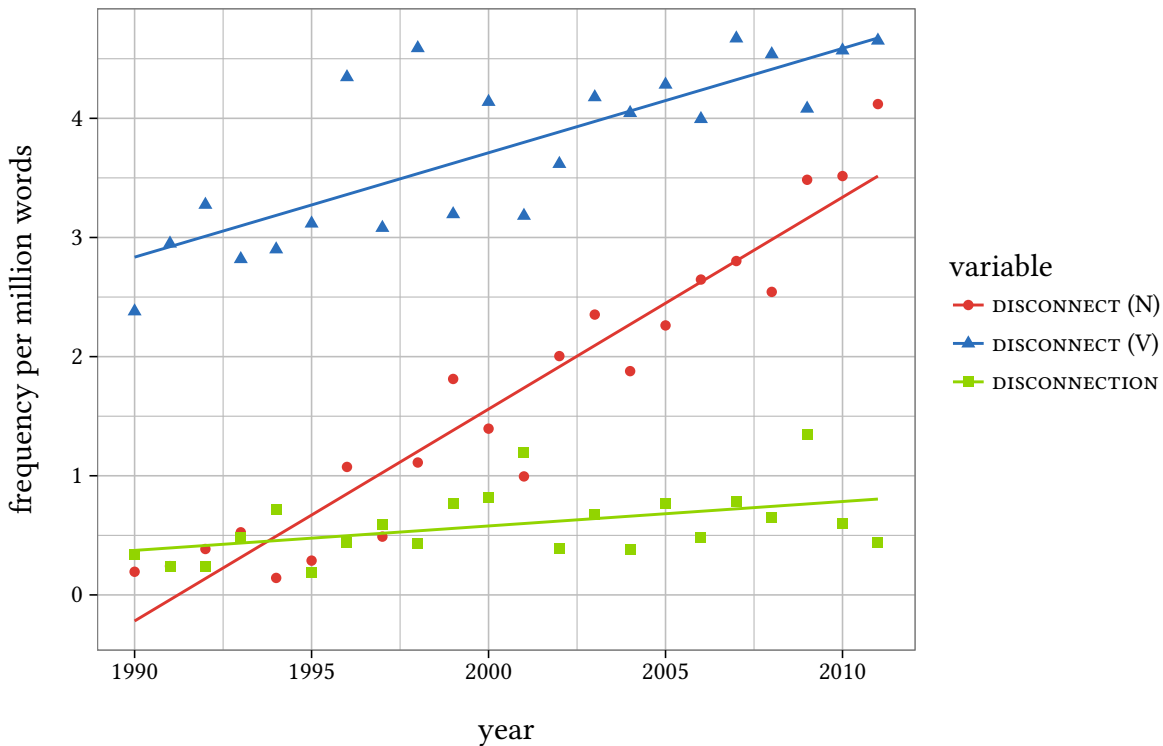


Figure 5.1: Scatter plot with linear regression lines for DISCONNECT (N), DISCONNECT (V), and DISCONNECTION

as can be expected in the case of blocking, where the rise of one nominal form comes at the expense of the other nominal form.

This generalized linear model predicts as the dependent variable the logarithmic odds of the realization of DISCONNECT as a noun vs. DISCONNECTION, the alternative based on derivation (cf. section 4.2.3 for an explanation of generalized linear models). As predictors, the year, the frequency of the verb and the frequency of the near-synonym are chosen. The model equation is shown in 5.2.⁸

$$\text{odds of converted form} \sim \text{year} + \log(\text{frequency of verb}) + \log(\text{frequency of deverbal noun}) \quad (5.2)$$

It has to be noted, however, that the frequency of the converted form is supposedly never completely independent of the frequency of the verbal form and the frequency of the

⁸The command in R is: `glm(cbind(frequencyN, frequencySyn) ~ I(year-1990) + log.frequencyV + log.frequencySyn, data = verb, family = "binomial")`. As with the linear model above, this model is comparatively simple, so that stepwise regression was not necessary to choose the final model.

alternative form.⁹ It is to be assumed that the frequency of all three forms will increase when the pragmatic need for them arises. As is shown below (cf. section 5.1.2), it is highly likely that the rise of digital technologies and the internet around the turn of the century has led to an increased need to talk about connecting to and disconnecting from devices and networks.

Table 5.2: Generalized linear model for DISCONNECT

	Estimate	Std. Error	z value	p	
(Intercept)	-0.22	0.27	-0.83	0.408	
year	0.10	0.02	4.31	0.000	***
verb	1.13	0.70	1.61	0.108	
deverbal noun	-0.88	0.20	-4.44	0.000	***

Null deviance: 102.93 on 21 degrees of freedom

Residual deviance: 16.50 on 18 degrees of freedom^a

^a Model fit greatly increases when including the independent variables. The null deviance is the deviance of the null model (no predictors). It decreases by 86.43 if predictors are included in the model.

The results for the generalized linear model are displayed in table 5.2. The model shows that the log odds of conversion increase highly significantly with every year, even though the estimated log odds are rather small ($B = 0.1$, $p < .001$). As expected, the frequency of the verbal form has no significant effect on the likelihood of conversion. In contrast, the frequency of the derived alternative proves to be a highly significant predictor for the odds of conversion ($p < .001$). The more frequent the derived form is, the less likely conversion becomes, as indicated by the minus sign (for an increase in the predictor variable by 1, the log odds for conversion change by -0.88). What the high significance of this effect testifies to is that the developments of the nominal forms are closely interwoven and influence each other. However, it cannot be said that the high token frequency of DISCONNECTION blocks (for blocking cf. section 2.1.2) nominal DISCONNECT, on the contrary: the relation is reciprocal. The rise of one form comes with the decline in usage of the other. Both forms can consequently be said to predict one another. Up until the mid nineties, both forms are equally infrequent in the corpus (cf. figure 5.1). While nominal DISCONNECT is firmly established as a noun by the year 2005, the use of DISCONNECTION develops at a much slower pace. From the early 2000s on, DISCONNECT can thus be assumed to be blocking DISCONNECTION. The high significance of the predictor ‘derived alternative noun’ can hence be interpreted as an indicator for the limited role of DISCONNECTION in preempting the spread of DISCONNECT.

⁹Potential collinearity in the data is addressed by centering the predictor frequency values, cf. section 4.2.5.

Statistical preemption in this case has rather reversed its direction with the suffixed form being preempted by the converted form.

CONNECT blocked

CONNECT, on the other hand, hardly ever occurs as a noun, is therefore not attested in the OED, and is thus a prime example of effective token blocking. The data are visualized in figure 5.2. In 1990, CONNECT as a verb is approximately 40 times as common as the nominal form. This consequently discourages the use of CONNECT as a noun, since the string “connect” is well entrenched as a verb in speakers’ minds (cf. Ungerer 2002: 560–563). Furthermore, the synonym CONNECTION is very frequent and occurs in a broad range of registers (cf. section 5.1.5). As is expected, high frequency of the blocking word preempts the spread of the new word. In this case, the blocking constraint is effective. The elevated frequencies of CONNECT (V) and CONNECTION thus inhibit the establishment and spread of CONNECT as a noun. In recent years, the frequency of use of CONNECTION has reached such heights (almost 80 times as often as CONNECT (N)) that it seems extremely unlikely that CONNECT (N) might spread in the future.

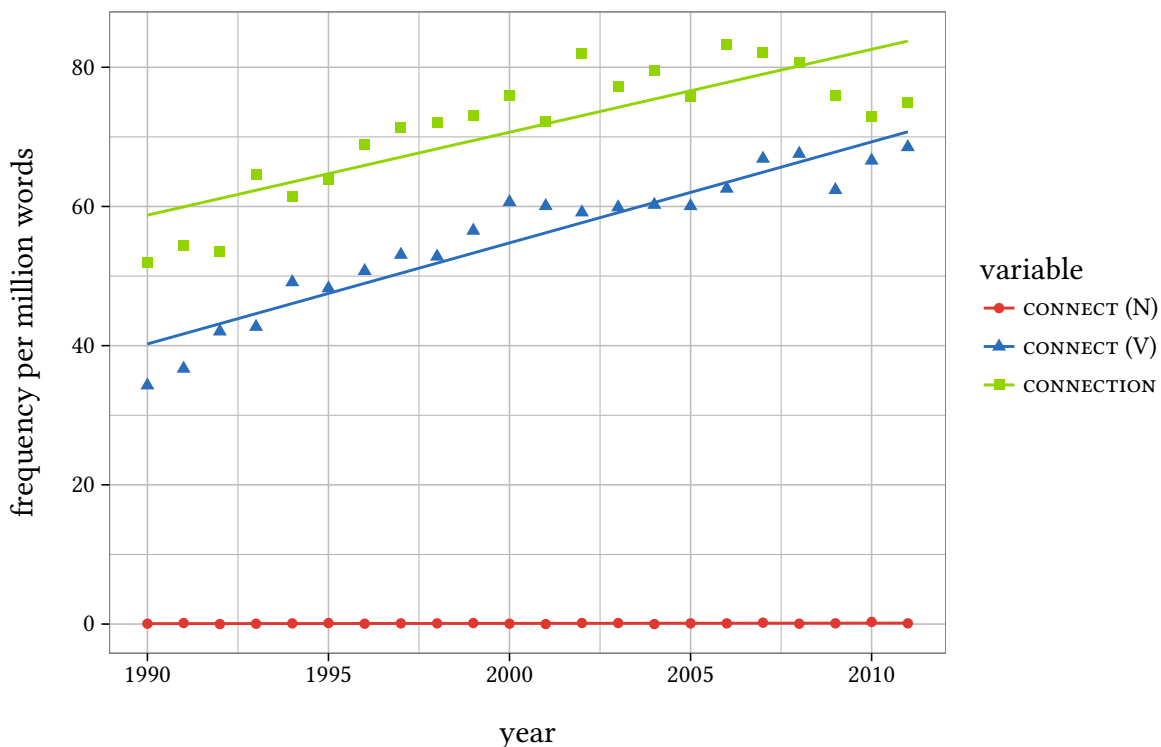


Figure 5.2: Scatter plot with linear regression lines for CONNECT

A linear model calculated on the basis of the actual frequency values is unnecessary here, as figure 5.2 clearly shows the developmental trends as well as the blocking effect which prevents the spread of CONNECT (N). Nevertheless, a linear model calculated on the basis of logarithmic values, the coefficients of which are shown in table 5.3, reveals that the relative increase in frequency over time is highest and highly significant for CONNECT (N), while it is not significant for the verb and the derived noun. Yet, the increase in relative frequency that this model with logged values shows is only minimal (0.03 increase in log frequency per million words for every year) and therefore does not change the overall picture for CONNECT. The increase in frequency of the nominal form is strongly blocked by the much more frequent occurrence of the verb and the blocking noun CONNECTION.

Table 5.3: Linear model for CONNECT

	Estimate	Std. Error	z value	p	
(Intercept)	−2.60	0.13	−20.60	0.000	***
verb	6.31	0.17	36.58	0.000	***
deverbal noun	6.67	0.17	38.70	0.000	***
year	0.03	0.01	2.68	0.010	**
verb : year	0.00	0.01	0.07	0.945	
deverbal noun : year	−0.01	0.01	−0.66	0.511	

5.1.2 Semantic shift

The general increase of the words DISCONNECT and CONNECT is most likely due to a higher communicative necessity of these lexemes caused by the rise of digital technologies such as computers and the internet, where connecting plays a central part. A collocation analysis of *connect*, *connects*, *disconnect*, and *disconnects* (span 4+4) in COCA reveals that among the most frequent collocates are indeed terms from the domain of digital technologies such as *internet*, *computer/s*, *cable/s*, or *network/s*.

Semantic changes which might have occurred between the years 1990 and 2012 are explored with the help of two wordles (cf. Feinberg 2013, figs. B.1a and B.1b in appendix B). Wordles make use of an algorithm which translates frequency into font size, that is, the larger the word appears in the wordle, the more frequently it occurs in the input text. One of the wordles is from 1990–1992 and the other from 2010–2012. Whereas the contexts of the earlier phase indicate that DISCONNECT is mostly used in specialized registers and specific contexts (corresponding to meaning 1.)—lexical items such as *phone*, *power*, *electricity*, *circuit* prevail—

the context of the tokens from the 21st century indicates a shift towards every-day vocabulary items such as *think*, *people*, *there's* and the like (potentially corresponding to meaning 2.).

Collocation analysis is drawn on to underline the intuitive findings from the wordle. The more sophisticated method reveals a significant semantic (and syntactic, see below) shift for DISCONNECT. In the following, the results for a collocation analysis of *disconnect* and *disconnects* in COCA are displayed. The span is 4 words in each direction; a word showing a mutual information score of 3.00 or higher is interpreted as a collocate of DISCONNECT. Only those collocates that occur at least five times have found their way into table 5.4.

Table 5.4: Collocation analysis for *disconnect/s* in COCA. MI scores, sorted by year and part-of-speech; absolute token frequencies given in parentheses.

collocate	1990–1994	1995–1999	2000–2004	2005–2009	2010–2012
prepositions					
<i>between</i>	(8) 3.60	(39) 5.08	(88) 5.73	(120) 5.90	(83) 6.09
there					
<i>there</i>				(88) 3.42	(57) 3.44
<i>theres</i>				(5) 5.70	
nouns					
<i>air</i>		(5) 3.48			
<i>button</i>				(5) 6.02	
<i>notice</i>	(5) 6.78				
<i>phone</i>	(5) 5.32	(5) 4.09			
<i>power</i>	(6) 4.14				
<i>reality</i>				(7) 4.73	(5) 4.98
adjectives					
<i>big</i>					(8) 3.04
<i>complete</i>				(5) 4.12	
<i>growing</i>				(5) 3.72	
<i>huge</i>				(8) 4.51	(5) 4.43
<i>major</i>				(6) 3.17	
<i>real</i>		(5) 3.21		(8) 3.13	(5) 3.13
<i>total</i>					(6) 4.36

The analysis shows that in 1990 DISCONNECT collocates with completely different words than in 2012. The only word that is a stable collocate of DISCONNECT is *between*. Nonetheless, the MI score indicates that the attraction between DISCONNECT and *between* has increased over time. Further collocates of DISCONNECT in 1990 to 1994 include *phone*, *notice*, and *power*. These nouns then disappear as collocates and in 2005 various adjectives, mostly adjectives

of dimension, take their place. Among them are *huge*, *major*, *complete*, and *growing*. Furthermore, in 2005, *there* appears as a collocate of DISCONNECT. Even though the collocation analysis does not allow for a differentiation between the adverb and *there* as the formal subject in the EXISTENTIAL construction, a glance at the corpus samples reveals that most tokens are instantiations of the EXISTENTIAL construction. The frequent occurrence of DISCONNECT in combination with *there* from 2005 on hints at the spread of the existential construction *there is a disconnect between*, which indeed rises significantly over the years (see below).

This collocation analysis provides support for the intuitive findings from the wordles. Over time, DISCONNECT has moved away from fairly technical vocabulary like *phone* or *power* towards semantically less loaded collocates, such as the existential *there* or the preposition *between*.

The following examples from the corpus illustrate the findings from the collocation analysis. Examples 5.3 and 5.4 indicate that in the early nineties, DISCONNECT was reserved for technical contexts, describing the action of cutting a connection. The examples from recent years (5.5 and 5.6) show adjectival collocates that are less semantically specialized and can occur in a broad range of contexts. The rising frequency of the existential construction with *there* also demonstrates that DISCONNECT has acquired a more general meaning but is still restricted to formulaic environments such as the partly schematic, partly substantive existential construction (see below).

- (5.3) The GFCI should indicate “open” or “off” and should disconnect the power from the GFCI-protected circuit. (MAG, 1992)
- (5.4) They couldn’t even disconnect his phone. (MAG, 1991)
- (5.5) And now theres [sic] a huge disconnect between his private life and his public persona. (SPOK, 2009)
- (5.6) So there’s a real disconnect between success and what your job is. (NEWS, 2007)

A similar semantic development can be observed for CONNECT. Table 5.5 displays the ten most frequent collocates for COCA and GloWbE. While the collocates before the year 2000 emphasize the physical act of connecting (to *computers* or *networks via cables* and *wires*) the second half of the table contains names of social networks such as *Facebook* or *Twitter*, that is, refers to the act of establishing a digital connection. Furthermore, a minor denotation (therefore not displayed in the table) that develops is the social aspect of establishing a connection, as is apparent from collocates such as *voters* (2000–2004, MI 3.47) or *audience* (2005–2009, MI 3.31). Generally, the semantic development of CONNECT also traces the path from physical,

fairly technical denotations (collocates include *cable* and *wire*) to broader, less semantically restricted meanings (e.g. *able*, *allows*, *opportunity*). Particularly the numerous collocates of CONNECT listed under nouns underline the important role of the lexeme in the context of new technologies (e.g. *internet*, *network*).¹⁰

5.1.3 Functional shift and syntactic variability

Functional shift as indicated by collocates

As far as the change in word class is concerned, the word classes of the collocates reveal the syntactic shift of DISCONNECT over time. During the first time span, 1990–1994, the group of collocates is almost entirely composed of nouns. Words that typically collocate with verbs are nouns, as can be concluded from syntactic patterns available in the English language. The simplest sentence construction, the SUBJECT-PREDICATE construction, is composed of at least a subject and a verb (SV), where DISCONNECT could serve as an intransitive verb, most likely in a construction such as the medio-passive, illustrated in 5.7.

(5.7) [T]hey all disconnect easily so that valuable electronics and equipment can be unfastened and stowed safely away. (MAG, 1992)

Furthermore, in the TRANSITIVE construction (SVO), the verb DISCONNECT takes an object which is realized by a noun phrase whose head is a noun. These patterns clearly show that nominal collocates in the case of *disconnect/s* very often imply a verbal node. Indeed, as has been shown above, in the years 1990 to 1994, DISCONNECT is mostly used as a verb. It occurs in contexts such as the following:

(5.8) The GFCI should indicate “open” or “off” and should disconnect the power from the GFCI-protected circuit. (MAG, 1992)

(5.9) They couldn’t even disconnect his phone. (MAG, 1991)

The only noun that DISCONNECT still collocates with in 2012 is *reality*. Closer inspection of the corpus tokens immediately reveals that this noun does not serve as the subject or object in sentences where DISCONNECT is used as a verb. It is rather the head of a prepositional complement in a prepositional phrase introduced by *between* that serves as a postmodifier of

¹⁰A few of the collocates merit commenting. First, *globus* is part of the name of a software. The name *Globus Connect* appears in one issue of one journal only. The collocate is listed for mere reasons of completeness. Second, the adjectival collocate *direct* could be interpreted as a premodifier of a potential converted noun *connect*. A look at the corpus data shows that this collocate is based on a new service called *Direct Connect* and is hence not an indicator of a tendency of CONNECT to convert.

5 Conversion as a productive process in US English

Table 5.5: Collocation analysis for *connect/s* in COCA and GloWbE. MI scores, sorted by year and corpus.

collocate	COCA					GloWbE
	1990–1994	1995–1999	2000–2004	2005–2009	2010–2012	2012
adjectives						
<i>able</i>			3.3	3.1	3.2	
<i>direct</i>			(4.2)			
adverbs						
<i>directly</i>	4.0	4.7	4.0	4.4	4.1	3.6
nouns						
<i>ability</i>	3.4	3.0	3.9	4.3		3.4
<i>bridge</i>	4.6				4.1	
<i>cable</i>	4.3	4.7	4.8			
<i>computer</i>	3.4	3.6	3.5	3.7		3.0
<i>computers</i>	4.9	4.2				
<i>devices</i>			4.8			
<i>dots</i>	7.4	8.8	9.1	9.0	9.4	9.5
<i>Facebook</i>					3.8	4.3
<i>globus</i>					(11.1)	
<i>internet</i>		4.6	4.5	4.2	3.6	3.5
<i>lines</i>				3.0		
<i>network</i>	3.6	3.6	3.9	3.5		
<i>networks</i>	5.3					4.0
<i>opportunities</i>					4.0	
<i>theory</i>				3.5		
<i>transit</i>						5.7
<i>Twitter</i>						3.9
<i>wire</i>	5.3					
<i>wires</i>		6.4				
prepositions						
<i>via</i>		5.1		4.5	4.3	3.8
verbs						
<i>allows</i>					4.1	

DISCONNECT (N). The meaning of this phrase is in most cases the difference between an ideal state and *reality*, as can be seen in examples 5.10 and 5.11.

(5.10) [This disconnect [between dream and reality]_{postmod}]NP characterized the Trustees' entire administration, however. (ACAD, 2009)

(5.11) “There’s [a disconnect [with reality]_{postmod}]NP” when it comes to literature for boys, Tripp says. (NEWS, 2011)

The prevalence of nouns as collocates of DISCONNECT quickly declines over the years following 1994 to give rise to a number of adjectives as collocates. Adjective phrases fill premodification slots in noun phrase constructions. The prevalence of adjectives as collocates consequently indicates that the conversion from verb to noun of DISCONNECT has happened or is on-going. This is illustrated in examples 5.12 to 5.14.

(5.12) And now theres [sic] a **huge** disconnect between his private life and his public persona. (SPOK, 2009)

(5.13) So there’s a **real** disconnect between success and what your job is. (NEWS, 2007)

(5.14) This is a **major** disconnect between our two cultures, and this is one of many problems with the war. (NEWS, 2007)

Table 5.6: Collocation analysis for *disconnect/s* in GloWbE (data from 2012). Absolute token frequencies given in parentheses.

collocate		MI
prepositions		
<i>between</i>	(541)	5.90
nouns		
<i>reality</i>	(54)	4.54
<i>internet</i>	(28)	3.34
<i>cable</i>	(13)	4.90
<i>battery</i>	(13)	4.87
adjectives		
<i>huge</i>	(33)	3.81
<i>growing</i>	(23)	3.81
<i>total</i>	(23)	3.31
<i>fundamental</i>	(22)	5.01
<i>complete</i>	(19)	3.25
<i>emotional</i>	(13)	3.86

An analysis of web-data from the year 2012 reveals further trends in the development of DISCONNECT. A collocation analysis of the presumably progressive web genres of the US section of GloWbE yields the results displayed in table 5.6 for the ten most frequent collocates. *Between* is, firstly, the most frequent and, secondly, the strongest collocate. The number of tokens is extraordinarily high compared to the other collocates. It occurs ten times more often than the second most frequent collocate. This is a strong indicator of the nominalization of DISCONNECT, as *between* serves as the preposition introducing prepositional phrases post-modifying nominal DISCONNECT. Furthermore, there are still a range of nominal collocates, three out of which (*internet, cable, and battery*) hint towards verbal uses of DISCONNECT. *Reality*, on the other hand, is once again mostly used in contexts such as the ones outlined above in 5.10 and 5.11. The rest of the ten most frequent collocates are adjectives. For *there* as a collocate, see below.

As far as CONNECT is concerned, table 5.5 offers a very clear picture. The collocates of CONNECT are mainly nouns, which indicates that CONNECT is overwhelmingly used as a verb. The adjectival collocates could hint at a potential nominalization. Nonetheless, *direct* proves to be irrelevant because it is part of the name of a service called *Digital Connect*, and *able* occurs in the [BE *able to* V_{INF}] construction, in which CONNECT fills the verbal slot. The analysis of the collocates of CONNECT thus tallies with the statistical analysis in that it does not indicate tendencies of conversion for CONNECT.

Functions in the clause and sentence patterns

As far as syntactic functions are concerned, DISCONNECT over time has come to fill many different slots in various different sentence patterns. While the TRANSITIVE construction is one of the most frequent constructions in which DISCONNECT (N) occurs, there are also instances of it filling slots in predicative constructions. The functional range of nominal DISCONNECT is illustrated by the following examples. It can occupy subject slots (5.15), object slots (5.16), complement slots (5.17 and 5.18), and can also be used in adverbials (A_{place} in 5.19, A_{time} in 5.20, A_{reason} in 5.21).¹¹

- (5.15) **The disconnect between Mr. Obama's public stance on lobbyists and his use of fund-raisers who are active in the lobbying industry rests in part on the ambiguity in the law over who must register as a federal lobbyist. (NEWS, 2011)**

¹¹The picture for CONNECT is fairly similar in that CONNECT (N) can occur in a range of syntactic functions. Yet, as with DISCONNECT, the TRANSITIVE construction prevails. Select examples are given in table B.1 in appendix B.

- (5.16) Phil was so surprised that he hardly registered **the complete disconnect between her action and her face, which showed no emotion at all.** (FIC, 2005)
- (5.17) And he was a **real disconnect** for me. (SPOK, 2001)
- (5.18) One of the problems cited by the Schlesinger report was **the disconnect between tactics authorized at Guantanamo, where “unlawful enemy combatants” were held and the Geneva Conventions did not apply, and the tactics authorized in Iraq where the president had said the Geneva Conventions did apply.** (MAG, 2004)
- (5.19) At the service panel, shut the power off at **the main disconnect** and remove the cover. (MAG, 1999)
- (5.20) Thanks to the constant drilling, Pride launched **within ninety seconds from disconnect and boost, ample time for a crew on hair-trigger readiness to strap into their battle stations.** (FIC, 2005)
- (5.21) But Professor Taylor and others contend that only very few will become active members, like Padilla, **in part because of an ideological disconnect.** (NEWS, 2002)

Chunks

The spread of DISCONNECT is due to several mechanisms, one of which is its embedding in frequently used constructions. As has been laid out in section 3.1.2, language is not perceived and processed word for word but rather in larger units, so-called chunks. The more frequently words co-occur, the more likely it is that they are stored as chunks, which can be shown by faster recognition times in experimental settings (cf. e.g. Arnon and Snider 2010: 76). In the production of language, chunks are produced more readily than their individual components because the entire unit is more easily accessible than its parts (cf. e.g. Janssen and Barber 2012: 10). Consequently, the embedding of a newly converted form in a frequently used construction will facilitate the spread of this new form. This diffusion via highly frequent constructions is another way of overriding the blocking constraint. It is thus not only the relatively higher token frequency of DISCONNECT compared to DISCONNECTION, but also its embedding in frequent constructions that has led to its spread over time. This mechanism is illustrated in the following for two constructions.¹²

¹²As it turns out, the LIGHT VERB construction is not relevant to the spread of *disconnect* (N). The converted form does not occur in this construction. This is not surprising when looking at the semantic constraints that Wierzbicka (1982: 758–759) and Dixon (2005: 469–470) posit for verbs that can occur in the LIGHT VERB construction. The verbs have to be “atelic” and “reiterative”; *disconnect*, however, can be argued to be neither, hence its non-occurrence in this particular construction.

DISCONNECT *between*. One construction in point is the NP construction with a fixed preposition in the postmodifying prepositional phrase. Collocation analysis has revealed that *between* is a very strong collocate of DISCONNECT. For convenience, the results from COCA and GloWbE are repeated in table 5.7.

Table 5.7: Collocation analysis of *between* and node DISCONNECT in COCA and GloWbE. MI scores, sorted by year and corpus.

	COCA					GloWbE
	1990–1994	1995–1999	2000–2004	2005–2009	2010–2012	2012
<i>between</i>	3.60	5.08	5.73	5.90	6.09	5.90

Between is already a collocate of DISCONNECT from the first years that COCA contains onwards. Over the course of time, the collocational strength as indicated by the MI score increases and is still high in the data from GloWbE. Minor differences in MI score numbers between COCA and GloWbE can most likely be attributed to the different registers that the corpora comprise. The numbers indicate that a high-frequency NP construction that is highly specific to DISCONNECT has emerged:

$$[\text{Det}_{\text{dtm}} \text{X}_{\text{premod}} \text{disconnect}_{\text{head}} [\textit{between} [\text{Y Z}]_{\text{NP}}]_{\text{PrepP, postmod}}]_{\text{NP}} \quad (5.22)$$

This construction, despite its complexity, is more substantive than the non-specific NP construction as shown in 5.28, which explains the high frequency of use of the former. The more substantive constructions are, the more readily constituents are identified and the more easily these constructions are processed. Easy processability leads to repeated use. This, in turn, results in a higher token frequency of the construction and a high token frequency then “leads to [the construction] being more cognitively entrenched” (Hoffmann 2014: 164, cf. section 3.1.2). Consequently, the frequent use of DISCONNECT in an at least partly substantive construction could possibly lead to a faster diffusion of this new form via this construction, particularly at the early stages.

EXISTENTIAL construction. A further construction that notably contributes to the spread of DISCONNECT as a noun is the EXISTENTIAL construction. *There* serves the function of dummy subject in the EXISTENTIAL construction [*there* BE NP], and DISCONNECT acts as the head of the noun phrase in the construction. In table 5.4, *there* appears as a collocate of DISCONNECT in the last two time frames. The MI scores for *there* as a collocate of DISCONNECT are repeated in table 5.8 for convenience.

Table 5.8: Collocation analysis of *there* and node DISCONNECT in COCA and GloWbE. MI scores, sorted by year and corpus.

	COCA					GloWbE
	1990–1994	1995–1999	2000–2004	2005–2009	2010–2012	2012
<i>there</i>	(2.12)	(2.58)	(2.81)	3.42	3.44	(2.59)

The years 1990 to 2004 show a steady rise in collocational strength, but the MI does not reach the critical threshold of 3.00 for *there* to be considered a collocate. This only happens in 2005 and *there* remains a collocate until 2012. The repeated use of the EXISTENTIAL construction and the fact that *there* has developed into a stable collocate are indicators for the increasing nominalization of DISCONNECT.

In contrast, in the data from GloWbE from the year 2012, *there* is not a collocate of DISCONNECT any more. This hints at more liberal uses of DISCONNECT in the web genres. In Construction Grammar terms, DISCONNECT is not restricted to specific constructions such as the EXISTENTIAL construction any more but can also be used to fill slots in other, even more schematic constructions. The following examples illustrate this for the TRANSITIVE construction.

- (5.23) Xylem’s survey reveals the **disconnect** between awareness over growing water scarcity and who they think should pay for the country’s water problems.
(GloWbE-US)
- (5.24) Of course, given that I am pointing out these **disconnects** in The New York Times, it will be seen as confirming what conservatives already know. (GloWbE-US)
- (5.25) I remember, several years ago, reading an essay that described one of the major **disconnects** between lit-fic and other genre fic (and the fans of the respective styles).
(GloWbE-US)

CONNECT (N). As regards CONNECT (N), the scarcity of nominal tokens does not allow for a quantitative analysis of potentially supporting constructions. Notwithstanding, a qualitative analysis of the corpus tokens reveals that CONNECT (N) occurs relatively frequently in both of the above-mentioned constructions. To some extent, this might be due to speakers modeling the use of CONNECT (N) in analogy to DISCONNECT (N), as examples 5.26 and 5.27 illustrate.

(5.26) CAVUTO: [...] Whats [sic] the disconnect? TOLL: Nobody ever said **there is a connect between** fortunes of company and the fortunes of the company stock. (SPOK, 2002)

(5.27) So we've got **this big connect – disconnect between** politics and economics; (SPOK, 2011)

5.1.4 Emergence of the plural form

Full conversion can be said to be achieved when the converted form has adopted all characteristics of the target word class. For nominalizations, an important formal characteristic is the existence of a plural form. It can be hypothesized that the establishment of the singular form will precede the emergence of a full nominal paradigm. The frequencies of the singular and plural forms of DISCONNECT are displayed in figure 5.3. The plural of DISCONNECT is attested in COCA from the year 1990 on, that is, already at a very early stage. Nevertheless, it is not widely used. Although a slight rise can be observed over the entire time span up until 2011, the plural form is infrequent compared to the singular form. The data from GloWbE show a similar imbalance of singular to plural form (1321 tokens of nominal *disconnect*, 69 tokens of nominal *disconnects*). It remains to be seen whether the plural form will increase in frequency in the same way that the singular form has increased.

Despite its comparatively low frequency, what becomes immediately evident upon studying the uses of the plural form in the corpus is that *disconnects* at a very early stage is already embedded in fairly complex noun phrase constructions, more complex than the singular form at the same stage of its development (operationalized by comparable frequency of occurrence).¹³ Once the singular form has 'paved the way', i.e. is well entrenched in certain constructional slots, the plural form is able to occupy these slots as well.

The NOUN PHRASE (NP) construction consists of four slots of which only one slot, the head, is obligatory:

$$[\text{Det}_{\text{dtn}} \text{X}_{\text{premod}} \text{N}_{\text{head}} \text{Y}_{\text{postmod}}]_{\text{NP}} \quad (5.28)$$

A determiner fills the determinative slot and a noun fills the head slot of the construction. The X slot of the premodification can be filled by an adjective phrase construction or another NP construction. The postmodifying Y slot can be filled by a prepositional phrase construction or by a relative clause construction. An NP construction with only the determinative and

¹³It must be acknowledged that due to data scarcity determining the point where the singular and the plural form have reached a similar developmental stage is almost impossible based on the data from COCA.

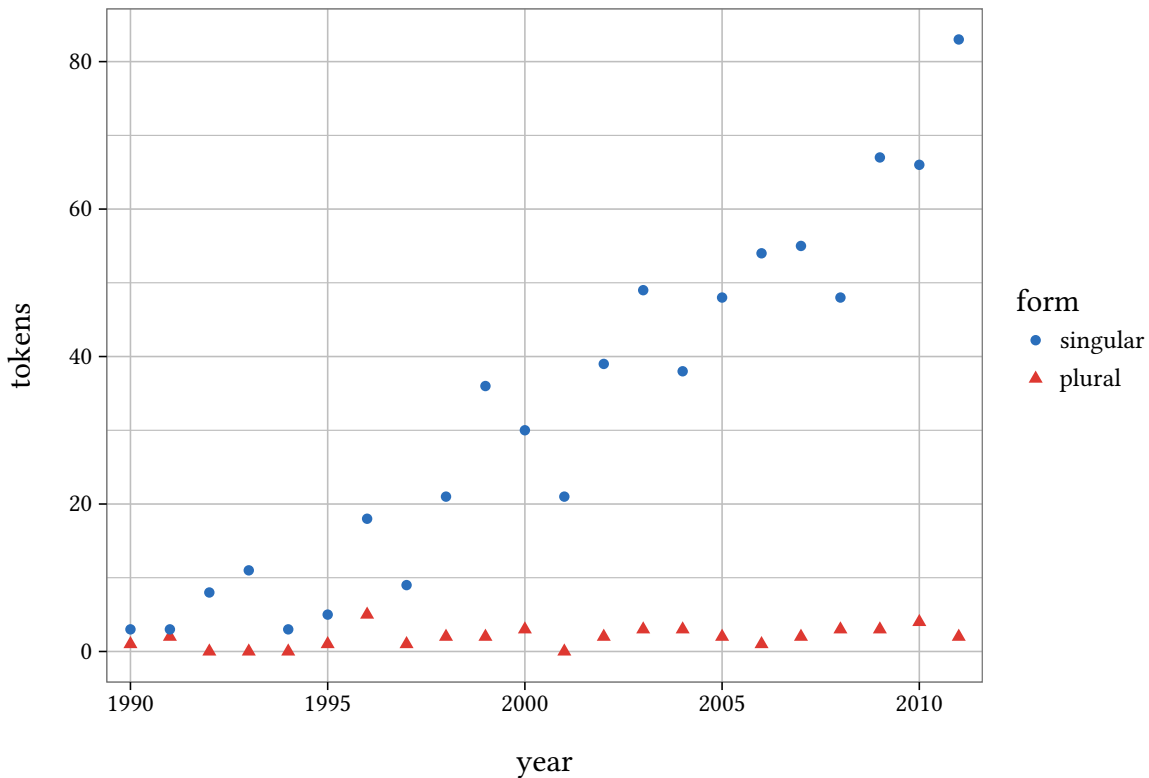


Figure 5.3: Frequency of singular and plural form of DISCONNECT (N) per year

the head slot filled can be considered less complex than an NP construction with the X and Y slots filled.

The first two years in COCA, 1990 and 1991, offer six tokens for the singular and three tokens for the plural form. While four of the uses of *disconnect* in the singular are embedded in minimally complex NP constructions, that is, simple determiner and *disconnect* as the head of the NP such as *a disconnect*, *this disconnect*, *no disconnect* (cf. examples 5.29 and 5.30), none of the occurrences of the plural is in such NPs. All instances of the plural appear in more complex constructions involving, for example, complex determiners like *a lot of* or *a series of*, adjectival premodification or clausal postmodification (cf. examples 5.31 and 5.32).

(5.29) He and other NPists claim that this disconnect is partly responsible for the popular rebellion against government. (MAG, 1991)

(5.30) I do envision, correctly, a disconnect between the leadership of Saddam Hussein and his armed forces. (SPOK, 1991)

(5.31) I think there's a lot of disconnects in Iraq, that Saddam is not fully briefed on everything that goes on. (SPOK, 1990)

(5.32) And AT&T, good old reliable 'Ma Bell', was 'in the soup' this year - a series of disconnects knocking out traders on Wall Street and airline passengers across the country. (SPOK, 1991)

Yet, despite the early embedding of the plural form in more complex constructions, it is also evident from examples 5.29 through 5.32 that the plural form 'lags' behind the singular form both in semantic as well as functional development. Example 5.32 illustrates the use of *disconnects* in its original meaning from the field of electricity, which prevails in the very early 1990s. In contrast, at the same time, the singular form is already predominantly used in its metaphorical meaning. Example 5.29 further shows that the singular *disconnect* in its metaphorical meaning occurs in subject position as early as 1991.

Another case which demonstrates the embedding of *disconnects* in complex constructions is its occurrence as the notional subject in the existential construction. Out of six tokens (1990–2012), only one shows the minimal form of a noun phrase, that is, $[\text{Det}_{\text{dtm}} \text{N}_{\text{head}} \text{Y}_{\text{postmod}}]_{\text{NP}}$,¹⁴ where the determinative and head slots are filled by a simple determiner and a noun only, respectively. All other tokens are embedded in more complex noun phrases that include complex determiners (*all these disconnects*) or premodifying adjectives (*no hidden disconnects, major disconnects*). There are further cases that present more complex realizations of the existential construction, e.g. with an adverbial inserted between the verb and the noun phrase (*there are **sometimes** striking disconnects that have an impact on the markets*, ACAD, 2002).

Once again, contrasting DISCONNECT with CONNECT provides interesting insights. As has been shown above, *disconnects* is established as the plural form of the converted noun, exhibiting a low yet steady frequency of occurrence. DISCONNECT can consequently be said to be an example of full conversion. The picture that CONNECT offers is very different. Out of 2436 tokens for *connects*, only 4 are nominal. The plural is thus a marginal phenomenon, so that CONNECT cannot be considered a case of full conversion.

5.1.5 Register analysis

Another aspect that merits detailed analysis is the registers in which DISCONNECT occurs. It can be assumed that conversion originates in registers closer to the conceptually spoken end

¹⁴In the case of DISCONNECT the postmodification can be considered obligatory if the meaning of the postmodification cannot be recovered from the context: *?I could see a disconnect*.

of the continuum and will spread from there to more formal, conceptually written registers, as is usually the case with linguistic innovations.

An analysis of the genres¹⁵ in which DISCONNECT vs. DISCONNECTION appear reveals that DISCONNECTION is to a large extent restricted to academic contexts (47%, cf. figure 5.4). It hardly appears in spoken discourse (8%). This void is filled by DISCONNECT (N). The newly converted lexeme is used in a broad range of genres (19% academic texts, 7% fictional texts, 16% popular magazines, 23% newspapers, 35% spoken register). It seems that what began as synonyms have evolved to become semantically differentiated words which show complementary usage patterns, with DISCONNECTION appearing mostly in academic and fictional texts and DISCONNECT covering the rest (magazines, newspapers, spoken text). What figure 5.4 further reveals is that with the exception of fictional texts all registers show a higher frequency of the converted form over the suffixed form.

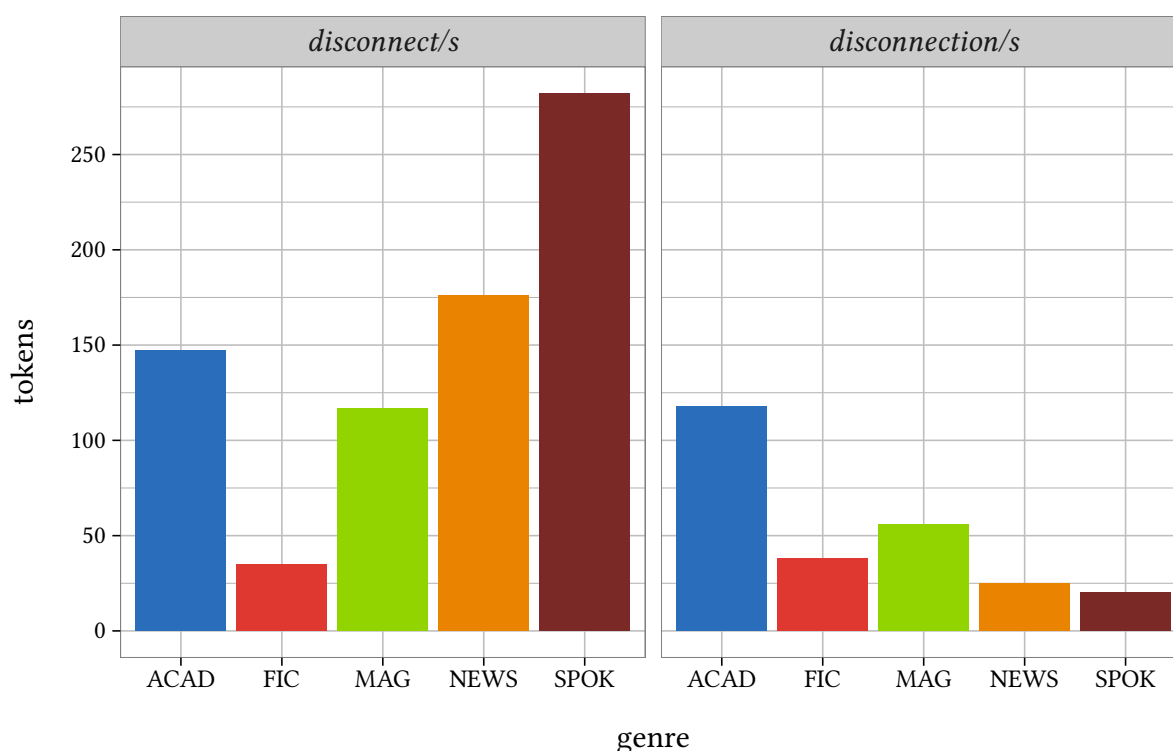


Figure 5.4: Distribution of genres for *disconnect/s* and *disconnection/s*

A year-by-year analysis of the registers (figure 5.5) illustrates how DISCONNECT (N) has evolved. From 1990 on, the spoken register is the dominant genre in which DISCONNECT (N) is used. However, particularly its usage in the news and the academic register is striking,

¹⁵Recall that the notions of *genre* and *register* as used here do not correspond to the strictly defined notions as proposed in textlinguistic studies such as Biber (1988), but refer to the sections of COCA.

5 Conversion as a productive process in US English

since the word hardly comes up until late in the 1990s just to then show a significant increase over the next ten years. Trend lines for every genre per year are displayed in figure B.2 in appendix B.

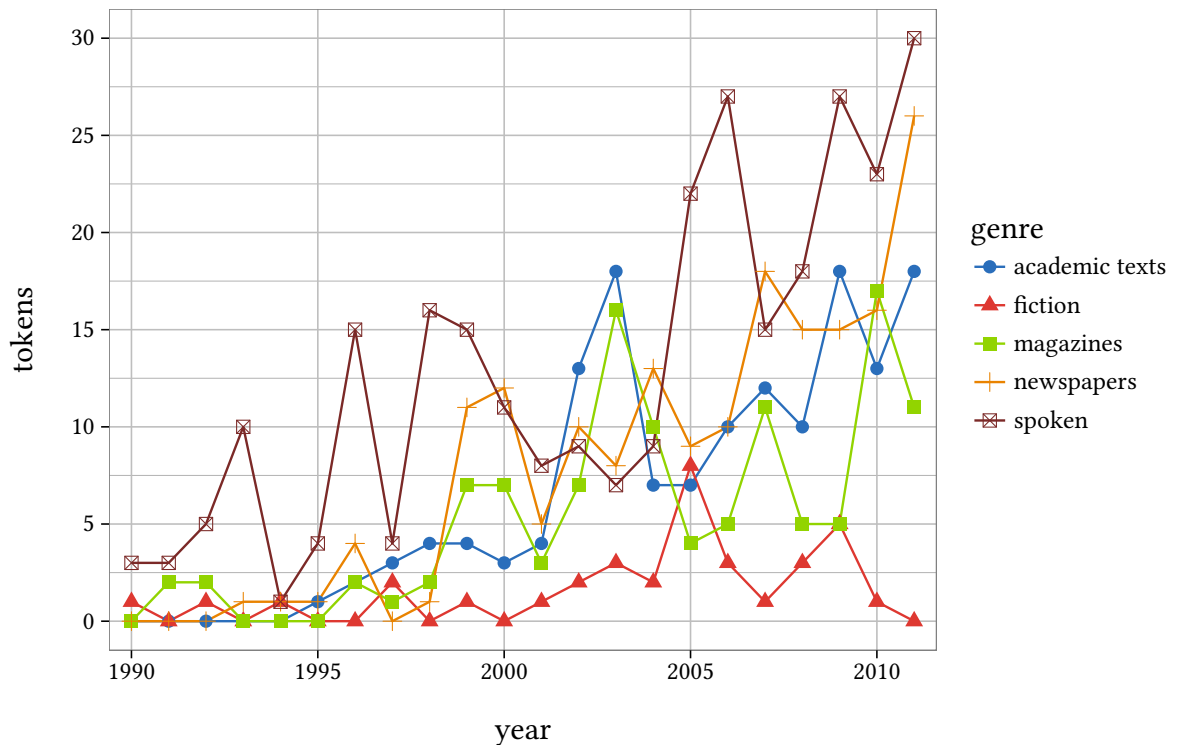


Figure 5.5: DISCONNECT (N) per genre per year

For DISCONNECTION (figure 5.6), its predominant use in the academic register across the years is immediately evident. This is to be expected since academic texts usually favor nominalizations and exhibit a comparatively higher number of longer words (cf. Biber 1989: 8, 12). DISCONNECTION is further used in magazines but hardly appears in the news or the spoken register. This is the case across all years. The analysis of the annual distribution by registers thus confirms the assumption that the converted and the suffixed form have taken on complementary functions, with the new, converted form being used more frequently in the spoken domain and the suffixed form pertaining mostly to the academic register. The converted form has occupied a void, the spoken register, and has subsequently spread from there to other registers, with the spoken and news register prevailing. That the news register should show similarities to the spoken genre is a result of the approximation of the news register to “oral styles” over the last century (Biber 2003: 170, also cf. Hundt and Mair 1999). Moreover, in order “to communicate as efficiently and economically as possible”, newspapers often favor a dense style marked by many short nouns (Biber 2003: 170). This could explain

why the news genre shows a preference for the converted form, which expresses the same concept but is shorter, over the suffixed form.

Nonetheless, it can further be observed that the converted form is also increasing in usage in the academic register at the same time that the derived form is used less in this register. It remains to be seen whether DISCONNECT (N) will take over more discursive functions from its rival DISCONNECTION.

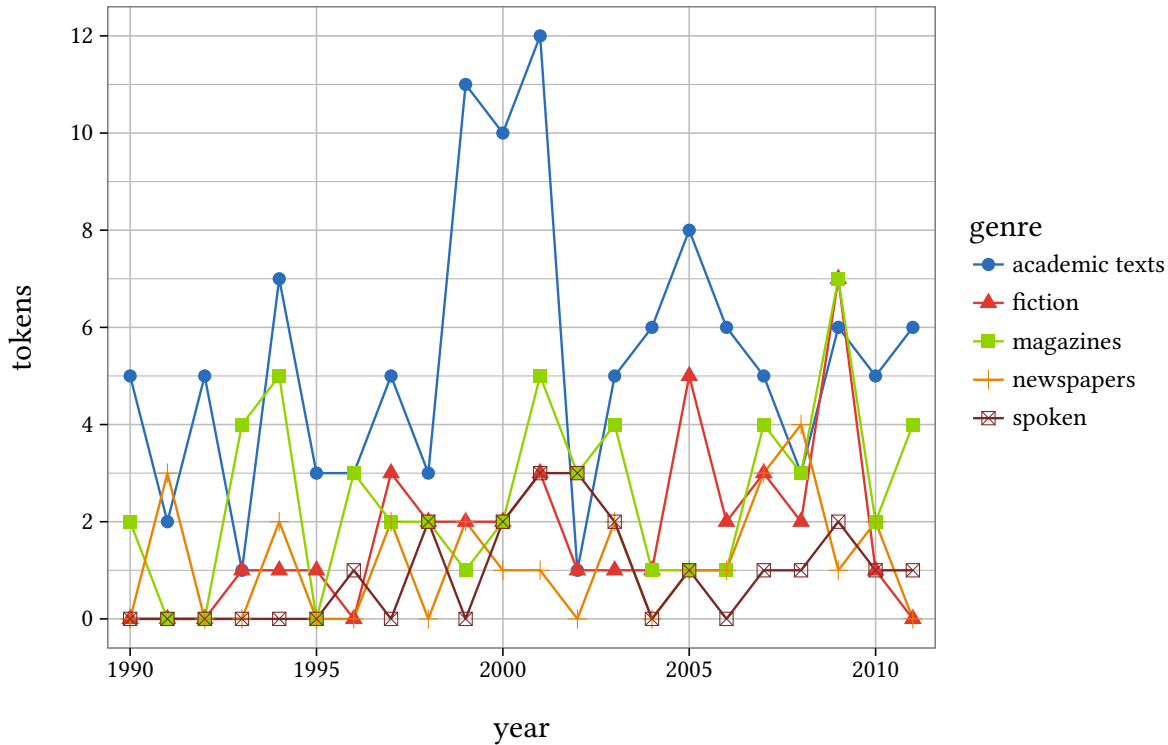


Figure 5.6: DISCONNECTION per genre per year

For CONNECT, the picture is very different. CONNECTION is well established in all registers (34% academic texts, 11% fictional texts, 23% popular magazines, 18% newspapers, 14% spoken register) and consequently preempts the spread of CONNECT in any of these contexts (cf. figure B.3 in appendix B). This is also in large parts due to the much higher token numbers of *connection/s* in all registers (cf. table B.2 in appendix B).

5.1.6 Interim summary

It can thus be concluded that the comparatively high frequency of CONNECTION constrains the spread of CONNECT (N), whereas the low frequency of DISCONNECTION and its restriction to mostly the academic register foster the establishment and spread of DISCONNECT (N). Furthermore, the fact that the word CONNECT is used much more often in contexts where it

serves as a verb leads to a higher entrenchment of CONNECT as a verb and thus hinders the change of word class. Since the noun-to-verb ratio of DISCONNECT is less skewed, a change of word class is facilitated, resulting in the increasing frequency of DISCONNECT (N).

The noun DISCONNECT over time acquires a metaphorical denotation that focuses on non-physical ways of non-functional connections. This is only possible once the nominal form has become independent of the verbal base. Additionally, DISCONNECT has developed a plural form that is also increasing in frequency (though slowly). Furthermore, over the years, DISCONNECT (N) has come to occupy diverse syntactic functions. DISCONNECT can hence be considered a prime example of successful verb-to-noun conversion, whereas the case of CONNECT illustrates the blocking mechanisms operative in the English language.

5.2 Further examples

The aim of this section is to show that the mechanisms in effect in the cases of DISCONNECT and CONNECT are not unique. This is illustrated drawing on further examples from US English. For these examples, random samples of size 100 were gathered for every single year between 1990 and 2012.^{16,17} Since COCA offers balanced data for every year, it is possible to analyze random data sets from each year.

5.2.1 DIVIDE

A case that illustrates an incipient process of conversion is DIVIDE. According to the OED (*OED Online* n.d., July 2015), the converted form has two meanings, one literal and one figurative, comparable to DISCONNECT (N). The literal meaning, “1. The act of dividing, division”, was first attested in 1642, whereas first attestation of the figurative meaning dates back to 1807. A quick glance at the data reveals that DIVIDE (N) is more polysemous than the entry in the OED suggests. The OED subsumes the use of DIVIDE in a geographical sense under figurative uses (“2. A ridge or line of high ground forming the division between two river valleys or systems; a watershed”). Also subsumed under this heading is the figurative meaning of DIVIDE as “a dividing or boundary line”. This classification of meanings can hardly be considered appropriate seeing that the metaphor is of a different quality in geographical expressions or proper nouns such as *the Continental Divide*, where a divide is easily visible

¹⁶The data for 2012 were excluded for the above-mentioned reasons.

¹⁷The values were then extrapolated to fit the corpus size. While the statistically versed reader might have reservations against this method, extrapolated values seem robust enough for the present analysis, in which trends rather than intricate detail in frequency of use are explored.

to the naked eye, compared to the more abstract meaning in formulations such as *the socioeconomic divide*, where the divide is only apparent on closer inspection of e.g. figures that document trends in society. For the present purposes, all uses of DIVIDE (N) as a proper noun (e.g. in *Continental Divide* or *Great Divide*, which refer to a specific hydrological divide on the North American continent, cf. Encyclopædia Britannica Online 2014) are excluded from the analysis, all other uses are included.

The linear model (for the formula cf. equation 5.1 on page 105) in table 5.9 shows that the nominal form rises in frequency, whereas the verb and the near-synonymous derivation, DIVISION, decrease significantly in frequency over time. The drop in frequency is more significant and more pronounced for the derived form ($B = -1.00$, $p < .001$) than for the verbal form ($B = -0.57$, $p < .05$). The scatter plot representing all the data points (figure 5.7) illustrates these numbers. While the verbal and the derived form decrease in frequency over the years, the nominal form DIVIDE is the only form that resists this general decrease in frequency of the lemma by showing a significant increase ($B = 0.39$, $p < .05$).

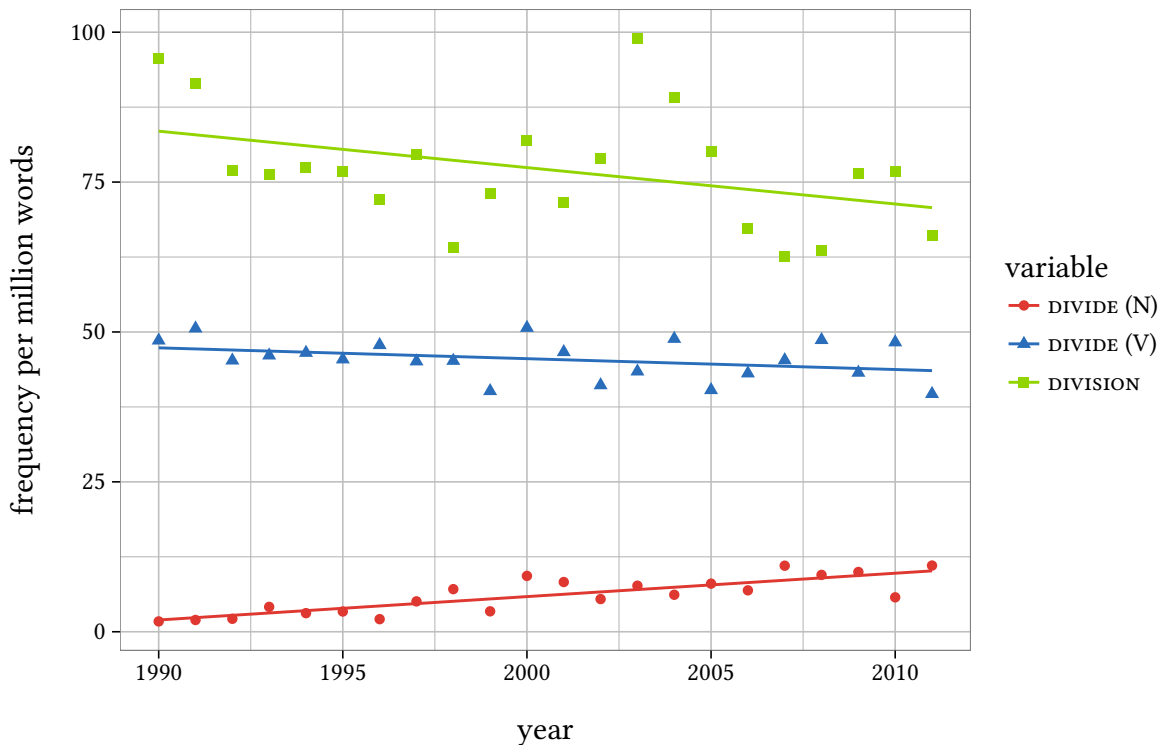


Figure 5.7: Scatter plot with linear regression lines for DIVIDE

When the same model is calculated with logarithmic values, it becomes even more obvious that the nominal form increases most in frequency. While the model with the actual frequencies calculates the absolute increase, the model with the logged values calculates the

Table 5.9: Linear model for DIVIDE

	Estimate	Std. Error	z value	p	
(Intercept)	1.96	2.38	0.82	0.413	
verb	45.39	3.36	13.49	0.000	***
deverbal noun	81.53	3.36	24.23	0.000	***
year	0.39	0.19	2.01	0.049	*
verb : year	-0.57	0.27	-2.08	0.042	*
deverbal noun : year	-1.00	0.27	-3.64	0.001	***

relative increase. In such a model, the increase for nominal *DIVIDE* is highly significant. Furthermore, the other forms differ highly significantly from nominal *DIVIDE* in their development in that their frequency decreases. (Cf. table C.1 in appendix C for the exact values and figure C.1 for a scatter plot of the corresponding values.)

As far as registers are concerned, *DIVIDE* is a good example of how a deverbal form is likely to establish itself. The number of tokens per genre in COCA are displayed in figure 5.8. As is evident from the graph, *DIVIDE* as a noun occurs in a range of genres in the first half of the observed time span, that is, up to the year 2000. After the turn of the century, the numbers for all genres except the academic keep increasing slowly yet steadily while the frequency of occurrence in the academic genre increases most rapidly in frequency. What these curves show is a differentiation in meaning that *DIVIDE* experiences. In the first years of use of the converted form, the form is not restricted to any specific contexts yet. This is apparent from the range of peaks in the first half of the plot. In the year 1993, it is most frequent in spoken discourse and magazines, in 1995, it is used mostly in the academic register. The year 1998 shows an increased usage in the spoken and fictional genres. From roughly 2000 on, the academic genre is the register in which nominal *DIVIDE* occurs most frequently. While all genres (except for fiction) record an increase in frequency over the entire time span due to the generally increasing use of *DIVIDE* as a noun, this development is most notable for the academic genre (cf. figure C.2 in appendix C).

Summing up, *DIVIDE* is similar to *DISCONNECT* in that the converted form is increasing in frequency despite the existence of a suffixed, near-synonymous form. This increase in usage frequency is moderated by the still higher frequencies of the verbal and the derived form, which are expected to slow down the spread of the converted form. Nonetheless, the data reveal that a semantic differentiation for *DIVIDE* is under way.

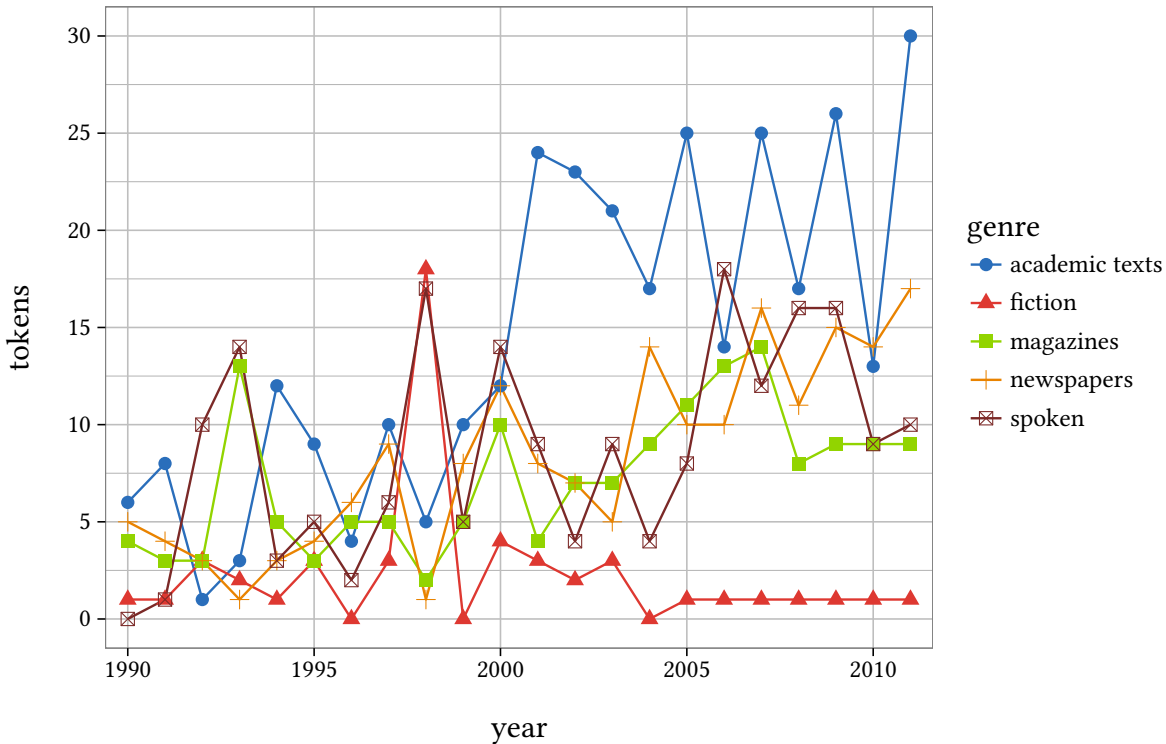


Figure 5.8: DIVIDE (N) per genre per year

5.2.2 INVITE

Another case of conversion is *INVITE*. The deverbal noun is near-synonymous to the derived form *INVITATION*. The first attestation of nominal *INVITE* denoting ‘the act of inviting; an invitation’ is from the year 1659. According to the OED (*OED Online* n.d., July 2015), *INVITE* (N) is marked as colloquial. This case resembles the case of *CONNECT* in that *INVITE* (N) is strongly blocked by *INVITATION*. For the random samples of size 100 for every year covering the time span between 1990 and 2012, COCA contains a total of 151 tokens of nominal *INVITE*. That is, out of 4670 analyzed tokens for *invite/s*¹⁸ only 151 were nominal uses. A quantitative analysis of such a small number of tokens is only of limited usefulness and will therefore not be performed.

Figure 5.9 shows the frequencies per million tokens for nominal and verbal *INVITE* and for *INVITATION*. A qualitative analysis of the token numbers shows a very slight increase in the frequency of nominal *INVITE* over time. Nevertheless, *INVITE* is without doubt a case where the blocking constraint is effective. The fact that the frequency of *INVITATION* is so much

¹⁸For *invites*, all occurrences, totaling 2370, were analyzed.

higher than the frequency of the converted form leads to a strong blocking effect: nominal INVITE is marginalized. What is more, the verbal form is considerably more frequent than the nominal form, which leads to the lemma being very well entrenched as a verb, as in the case of CONNECT. Considerably more processing effort than for the suffixed form would be required to correctly interpret (i.e. coerce) the converted form.¹⁹

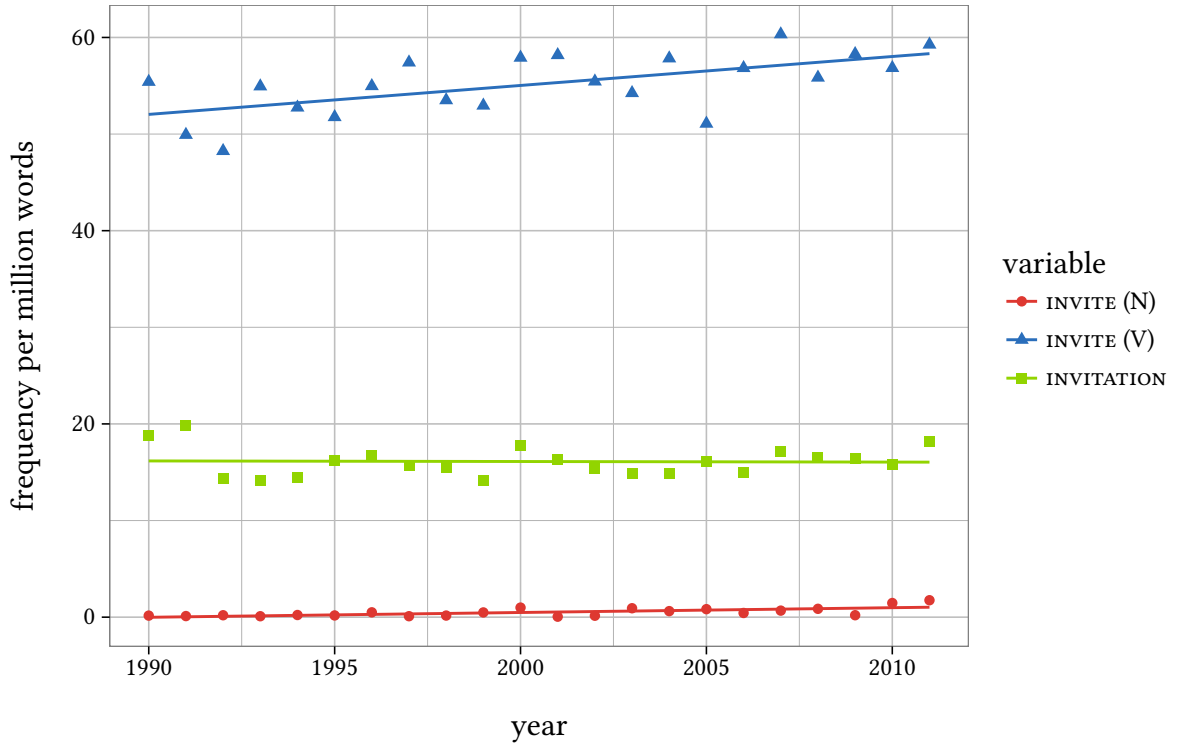


Figure 5.9: Scatter plot and trend lines for INVITE

Nevertheless, despite its marginal status, INVITE can be said to be an instance of full conversion. The plural form *invites* actually outnumbers the singular form and both the singular and plural form are used in a wide range of registers. As far as syntactic environments are concerned, the form fills subject and object slots alike. Nonetheless, the noun phrases in which *invite/s* fills the slot of the head tend to be of less than maximum complexity. Apart from the determinative slot, mostly only the premodification or the postmodification slot is filled, as examples 5.33 to 5.35 demonstrate.

(5.33) Natalie Taylor, 35, met some of the guys at a bar in the Oxford House last night and took them up on **an invite to come sit in**. (NEWS, 1994)

¹⁹As in the case of CONNECT, the logarithmic values as displayed in figure C.3 in appendix C reveal that the relative increase in frequency clearly is highest for nominal INVITE. This, however, cannot do away with the fact that both the verb and the derived noun are much more frequent and consequently block the emergence of the converted form.

(5.34) As a matter of fact, my first big adventure was that I scored **an invite to a Playboy mansion party**. (SPOK, 2000)

(5.35) It was a **pity invite**,²⁰ like when hunters bless a deer and then blow its head off. (FIC, 2000)

A notable exception is the following example which presents a noun phrase with all slots filled.

(5.36) Your BFF snags [[a]_{dtm} [coveted]_{premod} [invite]_{head} [to a New Year's Eve party]_{postmod}]_{NP} and takes you as her guest. (MAG, 2010)

5.2.3 PAY

Another example of conversion in American English is *PAY*, which is to a large extent in competition with the deverbal derivation *PAYMENT*. For *PAY* (N) the OED (*OED Online* n.d., July 2015) lists a number of meanings, indicating the polysemous nature of the word. The core meanings are “2. a. payment for a moral debt incurred; reward, recompense; [... ironic:] retaliation, punishment” (first attestation around 1300), as well as “3. a. [...] Money [...] paid for labour or service; wages, salary, stipend; remuneration” and “4. a. The action or fact of paying for something [...] As a count noun: a payment [...]” (first attestations around 1400 and 1440 respectively). The corresponding, closely related meanings of *PAYMENT* are numbers 4., 1., and 2. in the OED, and correspond to meanings 2. a., 3. a., and 4. a. of *PAY* (N), respectively. The other meanings mentioned in the entry for *PAY* (N) pertain to specific domains (e.g. military). Tokens with these meanings are excluded from the analysis. The entry furthermore lists a number of lexicalized compounds (e.g. *pay freeze*) that are also disregarded in the analysis.

Comparable to the situation for *CONNECT* and *INVITE*, the number of tokens of nominal *PAY* is too small to calculate meaningful regression models. Out of 4600 randomly selected tokens of *pay/s*, only 188 are nominal uses. A qualitative analysis of the scatter plot in figure 5.10 should suffice to explore the main trends. While the frequency of verbal *PAY* decreases over time, the frequencies for *PAY* (N) and *PAYMENT* remain level, indicating no effects for time on usage frequency. Moreover, the frequencies for *PAY* (N) and *PAYMENT* seem to be in a stable state of equilibrium with no evident blocking effect. A look at the plural forms, however, reveals a striking blocking effect. In the random sample there are only two occurrences of nominal *pays* (out of 2300 tokens). This means that in the corpus there must be considerably

²⁰*INVITE* also occurs in lexicalized compounds, as this example shows.

more tokens of the derived plural form *payments* compared to the converted plural form. It can therefore be hypothesized that *payments* heavily and very effectively blocks *pays* (N).

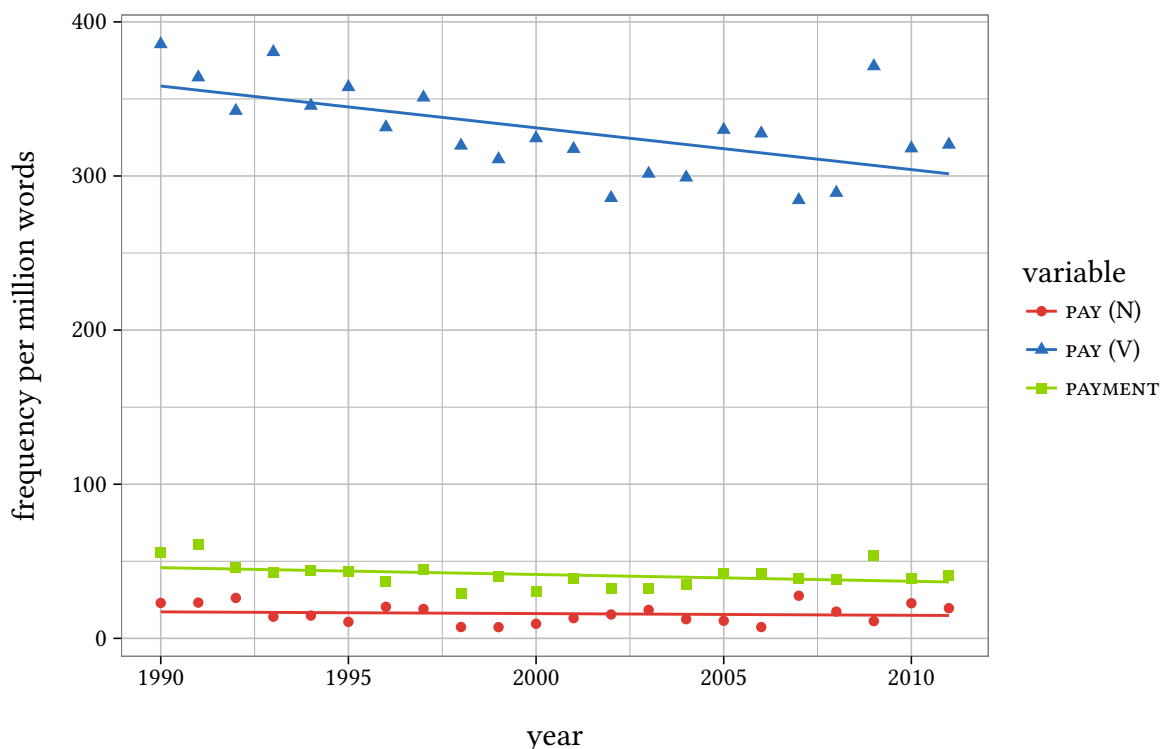


Figure 5.10: Scatter plot with linear regression line for PAY

Careful analysis of the corpus tokens further shows that *pay* is often used in compounds such as *pay r(a)ise*, *pay increase*, *pay freeze*, *merit pay*, or *overtime pay*. These compounds are highly lexicalized; they all have an entry in the OED (*OED Online* n.d., July 2015). In addition, even more lexicalized, more opaque compounds such as *pay phone* and *pay TV* (single stress only!) occur frequently in the corpus. Even though these tokens are excluded from the analysis, they impact speakers' mental representations of *pay*, which can lead to a preference for *pay* in compounds and a subsequent increase in frequency in these constructions as well as a decrease in other constructions (such as the DEVERBAL CONVERTED NOUN construction).

5.3 Summary: Conversion in USE

What this chapter has illustrated is that verb-to-noun conversion is a productive process in American English which yields nouns such as *DISCONNECT* and *DIVIDE*. Nevertheless, the success of conversion depends on various factors, one of which is the frequency of the competing derived noun. This effect has become known as the blocking constraint and has

been shown to be operative in cases such as *CONNECT* and *INVITE*, where the converted noun is marginalized due to a much higher frequency of occurrence of the competing, i.e. blocking, lexeme.

A further factor which has been identified is the frequency of the basis, that is, the verbal form. If the verb is very frequent compared to the converted form, then a change of word class is unlikely, owing to the firm entrenchment of the word as a verb. When “[o]nly the source item of the conversion [i.e. the verb] is well entrenched”, “[t]he processing effort required [to interpret the novel formation] is substantial”, as Ungerer (2002: 561) notes. In order to minimize the cognitive burden, speakers prefer not to convert the verb but use the suffixed/derived noun. This preference for the derived form, however, will only hold in instances where the derived form is sufficiently frequent to be so well entrenched and easily accessible that the processing of it is much faster than that of a novel converted form.

Another factor is the embedding of the newly converted form in frequently recurring chunks. The detailed analysis of *DISCONNECT* has demonstrated that its repeated use in the *EXISTENTIAL* construction, particularly with following *between*, plays an important role in the diffusion of the converted noun.

The success of conversion can be interpreted in terms of mere frequency but also in regard to how fully converted the form is. Apart from *DISCONNECT*, *INVITE* has proven to be a case of full conversion, even though its usage frequency is minimal compared to the blocking *INVITATION*.

Finally, it can be hypothesized that conscious speaker choices contribute to the spread of a converted form. It seems that some text types are more prone to creating or taking up newly converted forms. Among these figure the magazine and the news genre in COCA. Magazines might draw on newly coined forms to appear ‘trendy’, that is, to stress that they are following the latest fashion, even linguistically. The women’s fashion magazine *Cosmopolitan*, for example, uses *INVITE* (N) quite frequently; 12 out of 151 nominal tokens stem from this source. The news genre might also embrace converted forms, since they are shorter than synonymous suffixed forms and hence condense information.

After providing an in-depth analysis of verb-to-noun conversion in USE, a native variety of English, the next chapters will explore how the same phenomenon plays out in Asian varieties of English compared to USE and BrE, thus adding the dimension of language contact. Once again, a large corpus serves as the database. A quantitative analysis is presented in chapter 6, while chapter 7 offers a qualitative analysis of the corpus data.

6 A quantitative approach to conversion in World Englishes

Verb-to-noun conversion is a feature which is expected to lead to gradual rather than categorical differences between varieties, as has been pointed out above. That is, it is highly likely that varieties will develop statistically distinctive usage patterns for verb-to-noun conversion. Detecting those usage patterns consequently requires large amounts of data gathered from corpora that considerably exceed the size of the ICE corpora. All data in this chapter are therefore drawn from the *Corpus of Global Web-based English*, the smallest section of which comprises approximately 40 million words (cf. section 4.1.3). After presenting the hypotheses for this quantitative approach to V>N conversion, the data sampling procedure is described. The remaining part of the chapter deals with the results of various logistic regression models.

6.1 Hypotheses

Hypothesis 1: The frequency of conversion differs in varieties of English

The first hypothesis that guides this study is that varieties of English are expected to differ with regard to the frequency of use of V>N conversion. This is expected to be due to two phenomena. The first is transfer from the substratum, the second is the socio-institutional status of English in the respective region. Substrate transfer in V>N conversion is estimated to manifest itself in a different productivity of conversion, operationalized by frequency of use. That is, the more frequent V>N conversion is, the more productive the process is assumed to be in the respective variety. Following Bao's (2009) idea that structural convergence or divergence between the systems of the sub- and superstratum determines the productivity of constructions (cf. section 2.4), it is projected that productivity is a direct function of substrate influence, with convergent structures showing high(er) levels of productivity. The impact of the grammatical system of the contact language is expected to be particularly profound for

varieties with a substratum favoring non-morphemic word-formation processes, i.e. HKE or SgE, where Chinese is the main contact language. As a language that knows (almost) no derivation, Chinese is expected to foster the morphologically simple process of conversion. In order to test whether effects observed for HKE and SgE can be traced back to the analytic substratum, a third variety of English from the Asian context but with a synthetic substratum is introduced as a reference. Within the circle of Asian varieties, Indian English is an appropriate choice. It shares British English as the parent variety with HKE and SgE. Furthermore, Indian English emerged as a contact variety of English with Hindi, Bengali, and Telugu (among others), which are all synthetic languages (cf. section 1.1.3). Were the Chinese substrate the sole reason for variation between HKE and SgE on the one hand and BrE on the other, then IndE should not show an increased productivity of verb-to-noun conversion, since the synthetic Hindi substratum is unlikely to foster conversion in the same way as Chinese (cf. section 2.2).

The second explanation for distinct usage patterns of V>N conversion is the degree of institutionalization of the English language as laid out in the Dynamic Model (cf. Schneider 2007, cf. section 1.1.3). IndE and SgE share the same socio-institutional status in the Dynamic Model; both varieties are at stage 4, endonormative stabilization (with IndE arriving at stage 4 and SgE on its way to stage 5, cf. Mukherjee 2007: 170; Schneider 2007: 171). Were the degree of institutionalization the sole factor determining the success of verb-to-noun conversion, then SgE and IndE should show similar usage patterns of V>N conversion. The native varieties (USE and BrE) and HKE, as the most and the least established varieties of English, should exhibit opposing usage patterns. If HKE presented high numbers of V>N conversion, BrE (and USE) should display a considerably weaker inclination for V>N conversion.

Hypothesis 2: The blocking constraint is a global phenomenon

The second hypothesis is that the blocking constraint as laid out for US English in chapter 5 also applies to other varieties of English. While it is highly likely that the above-mentioned tension between transfer from the substratum and socio-institutional status of English has an effect on how strong the blocking constraint is, it is plausible to assume that the speaker's general tendency for economic language use—and with it, the avoidance of near-synonyms, i.e. blocking—is a world-wide phenomenon.

Hypothesis 3: The higher the usage frequency of a verb, the less likely conversion becomes

Thirdly, as far as intra-varietal variation is concerned, the aim of this study is to determine whether high- and low-frequency verbs differ as regards their likelihood to be converted to nouns. That items of high and low frequency are processed differently by the speaker is the key idea of the usage-based paradigm. It is generally assumed (cf. e.g. Bybee 2010) that more frequent items are more easily accessible than less frequent ones, since the speaker has more experience with the former. Following numerous studies rooted in the usage-based paradigm (cf. e.g. *ibid.*), it is hypothesized that a difference in frequency will lead to a difference in language processing which will manifest itself in a skewed pattern of productivity of verb-to-noun conversion.

The assumption is that a higher frequency of the base reduces the odds of V>N conversion. Since frequent verbs are strongly entrenched as belonging to the word class ‘verb’, a use of these forms in different word classes is rather unlikely, as has been shown for *CONNECT* in USE (cf. chapter 5). Along these lines, Teddiman (2012) found that in an experimental setting ambiguous forms that can be used as either verbs or nouns were mostly characterized as belonging to the word class in which they are most frequently used. That forms of a high usage frequency show “increased morphological stability” and consequently resist change to a greater degree than less frequent elements is known as the conserving effect of frequency (Bybee 2010: 24–25).

6.2 Corpus samples

For the quantitative study, a number of verb-noun pairs modeled on *disconnect* – *disconnection* were selected from two frequency classes (high and low). The frequency class of the verbs was determined drawing on a concordance list of all verbs occurring in ICE-Hong Kong. Those verbs out of the one hundred most frequent verbs in ICE-Hong Kong that have corresponding, derived nominal counterparts (*improve* > *improvement*) that do not denominate the person performing the action were determined. Verbs that only have derived nouns ending in *-ing* were excluded because of the possible confusion between present participle and the singular form of the noun (e.g. *understand* > *understanding*). Furthermore, verbs and nouns that only differ minimally in writing and/or sound (minimal pairs) were rejected, e.g. *believe/belief*, *live/life*, since they pose a high risk of accidental spelling mistakes. Deverbal nouns that are the product of multiple derivations (*learn* > *learnability*) were also discarded,

as multiple derivations can be assumed to be processed differently than single derivations (cf. Pliatsikas et al. 2014: 52). In a second step, all verb forms that have already been converted (e.g. *remains*, *estimate*) or can also denote an adjective (e.g. *direct*) were excluded. For this procedure, the OED (online edition) served as a reference.¹ Converted forms that according to the OED are obsolete were included. This procedure yielded a group of 18 high-frequency verbs and a group of 28 low-frequency verbs.

Random samples from GloWbE of size 1000² were drawn for the infinitive and 3rd person singular forms of twenty randomly selected verbs out of the two above-mentioned groups (ten each, see table 6.1).

Table 6.1: Randomly selected verbs and corresponding deverbal nouns

high frequency		low frequency	
verb	deverbal noun	verb	deverbal noun
allow	allowance	approve	approval
choose	choice	calculate	calculation
consider	consideration	deny	denial
continue	continuation	distribute	distribution
create	creation	examine	examination
develop	development	expand	expansion
improve	improvement	imagine	imagination
provide	provision	possess	possession
refer	reference	satisfy	satisfaction
require	requirement	specify	specification

The samples were then coded for part-of-speech (POS), both automatically (with the help of a computer script) and manually. Frequencies of verbs and nouns in the entire corpus were

¹Another option, beside a dictionary as a reference, could be the use of a major corpus of a native variety of English (or frequency data obtained on the basis of such a corpus). As verb-to-noun conversion is comparatively infrequent, the corpus would need to be large. Corpora that come to mind are the BNC or COCA. Additionally, the corpus should contain recent data, which excludes the BNC. Yet, this would mean comparing COCA with itself in chapter 5, as well as comparing GloWbE to COCA in the present chapter. As this would imply running the risk of a circular argument in the first case and comparing two very different corpora (cf. section 4.1) in the second case, the OED as a major dictionary of the English language is chosen as a reference. Seeing that dictionaries are edited, they can be considered a fairly objective record of the language. Yet, the editing process is time-consuming, which means that dictionaries are generally slower in documenting current language use than recently compiled or updated corpora (such as GloWbE or COCA). Also, dictionaries, by their very nature, are more conservative than actual language use. Nonetheless, in the present context, the OED appears to be the best option. In order to obtain the latest version, the online edition is drawn on.

²The samples were smaller when the form yielded less than 1000 hits in GloWbE.

then extrapolated from the resulting counts.³ The data were further coded for variety, and the token frequencies for the corresponding deverbal nouns formed by derivation were extracted from GloWbE. For mathematical reasons, frequencies were centered and logarithmically transformed. In total, 160,357 tokens, corresponding to 20 different verbs, were coded for part-of-speech. Out of those, 329 tokens were classified as nouns and 159,413 as verbs. 615 tokens could not be classified as belonging to either of the categories and were coded as NA. The extrapolation yielded a total of 1907 nominal tokens. Table 6.2 gives an insight into what the coded data look like.

6.3 Results and discussion

6.3.1 A ‘colonial’ model of conversion in Englishes

The predictor variables in the logistic regression analysis include the variety (GB, HK, IN, SG), the frequency of the verb (*frequencyVerb*)⁴, and the frequency of the corresponding deverbal noun (*frequencyDeverbal*). British English is set as the reference level since it constitutes the parent variety of all other varieties. The dependent variable is the nominalization of a verb through conversion or derivation calculated on the basis of the number of nominal tokens (e.g. *require/s*) and the number of tokens of the derived form (e.g. *requirement/s*). The odds of realization of a converted or a derived form are calculated based on the variety of English, the frequency of the deverbal noun, the frequency of the verb and the interaction between these variables. That is, the frequency of the deverbal noun and the verb are considered separately for each variety. Individual verbs are included as random effects, that is, the fitting of the statistical model is performed in such a way that effects will not depend on individual verbs. Equation 6.1 summarizes what the model calculates. Subjecting the model to an analysis of

³The use of extrapolated values as input for a logistic regression model is unusual. Through the extrapolation, effects in the counts that result from the random sampling procedure might be multiplied. This could distort the picture that the logistic regression model presents. However, POS-tagging all instances of all verbs would have been unfeasible and would necessarily have come with a restriction of the set of verbs to be analyzed. The method of extrapolation was consequently chosen for the sake of a larger set of lexemes to be studied. The size of the random samples, 1000, was large enough to cover over 10% of all occurrences of the potential singular for 68% of all verbs. For 24% of all verbs, the sample covered more than 50% of all occurrences. Of the potential plural form, over 50% of all occurrences of 62% of all verbs were covered. The counts can thus be assumed to be fairly reliable for the potential singular form and highly reliable for the potential plural form.

⁴Participles were not included in the calculation of the total frequency of the verbs. While GloWbE allows to search for lemmatized forms, thus excluding phenomena such as marginal prepositions (cf. Quirk et al. 1985: 660, 667–668), the overall quality of the tagging is not very reliable (cf. section 4.1.3). Particularly the past participle, a form that can be found in verbal and adjectival contexts, is prone to be mistagged. Consequently, participles were excluded from the frequency count.

Table 6.2: Subset of the coded data

id	query	variety	corpus	genre	source	token	POS
1	approve	HK	glowbe	HK G	inruld.org	and such experiences would provide useful lessons for other developing countries. # After the approve of the project of Experience of Universalizing the Nine-year Compulsory Education in Rural Areas in 's claim. # The other options are Follow The Wind (Angland) and Approve (Callan). # In the middle pin, follow Whyte's lead and a possible country way out with a small indefinite amount little bump of winning it approve. Nevertheless, do not get laid enough weapons system and fixtures to put up to guarantee the process capacity and quality. 6. PPAP (? Production Part Approve Process) lets the customers to see the quality and gets approval from the customers	n
2	approve	HK	glowbe	HK G	racing.scmp.com	body fit as well as cucumbers. Top the actual seafood cravings. # Single approve of good complete belly exercise program; # A supplements can help counterproductive in building	NA
3	approve	HK	glowbe	HK B	ziselbride.com	# 5. # other person as the Executive Committee may from time to time approve for admission, by invitation or otherwise. # have excelled in the promotion of	v
4	approve	HK	glowbe	HK G	kousheng.com	this (noun) sometimes as "right" (adjective) or "approve" (verb). Some translators, following Graham, translate fei as	v
5	approve	HK	glowbe	HK B	bjizwulu.com	Hong Kong # order police investigations of death # order inquests to be held # approve removal and use of body parts of the dead body # issue certificates of fact	v
6	approve	HK	glowbe	HK G	seatransport.org	agency is not yet using that power. # The same law lets the agency approve new tobacco products that could be marketed as safer than what's currently for sale	v
7	approve	HK	glowbe	HK G	philosophy.hku.hk	one of the EU member states, in principle, other member states should also approve the drugs to sale in the countries. But enterprises still need to get MA	v
8	approve	HK	glowbe	HK G	judiciary.gov.hk		
9	approve	HK	glowbe	HK B	...o.clearthear.org.hk		
10	approve	HK	glowbe	HK G	sinopharm.com		

the prediction error reveals that the model predicts with a higher accuracy than the baseline model (cf. section 4.2.6).⁵ The results of the logistic regression model are reported in table 6.3, additional coefficients are given in table D.1 in appendix D.1.⁶

$$\begin{aligned}
 \text{odds of converted form} &\sim \text{variety} + \text{frequency of deverbal noun} + \text{frequency of verb} \\
 &+ \text{variety} : \text{frequency of deverbal noun} \\
 &+ \text{variety} : \text{frequency of verb} \\
 &+ (1 \mid \text{verb})
 \end{aligned}
 \tag{6.1}$$

Table 6.3: Conversion in British vs. Asian Englishes

	Estimate	Std. Error	z value	p	
intercept (Intercept)	-6.93	0.27	-26.08	0.000	***
varieties					
HK	1.77	0.12	14.76	0.000	***
IN	0.76	0.11	6.96	0.000	***
SG	0.81	0.12	6.62	0.000	***
frequency of deverbal noun frequencyDeverbal	-0.43	0.14	-2.99	0.003	**
frequency of verb frequencyVerb	0.05	0.22	0.25	0.804	
variety : frequency of deverbal noun					
HK : frequencyDeverbal	-0.10	0.08	-1.27	0.204	
IN : frequencyDeverbal	0.24	0.08	2.82	0.005	**
SG : frequencyDeverbal	-0.07	0.10	-0.68	0.498	
variety : frequency of verb					
HK : frequencyVerb	-0.32	0.08	-4.12	0.000	***
IN : frequencyVerb	-0.14	0.08	-1.69	0.090	.
SG : frequencyVerb	-0.18	0.11	-1.64	0.102	

Conversion is more successful in new varieties

All new varieties differ significantly from BrE in that they show a higher chance of V>N conversion (as indicated by the estimates in the block ‘varieties’). The largest difference can

⁵The prediction error for the baseline model is 0.162%, while for the logistic regression model it is 0.082%.

⁶Since all predictors and interactions between them turned out to be significant, the original model was not modified by means of stepwise regression.

be found between HKE and BrE, with the odds of V>N conversion being 5.88 times as high⁷ in HKE. SgE and IndE differ only slightly from each other, with the odds of V>N conversion being 2.25 and 2.13 times as high as in BrE, respectively. It is highly likely that these two varieties do not differ significantly from one another with regard to V>N conversion.⁸

These results show that language contact does not suffice as the only explanation of contact variety grammar. First, all new varieties show a higher success of verb-to-noun conversion, which means that differences between the native and new varieties cannot exclusively be traced back to substratal influence. If V>N conversion were fostered only by a Sinitic substratum, IndE should not show a greater likelihood of conversion than BrE. There must hence be another mechanism at work that influences the productivity of conversion in the contact varieties.

HKE and SgE show different patterns despite a shared substratum

Second, HKE and SgE differ in how successful V>N conversion is. (Once again, a direct comparison is not possible, yet the estimates indicate considerable differences between HKE and SgE.) Verb-to-noun conversion is much more frequent in HKE than in SgE despite the fact that both varieties share a substratum that favors V>N conversion. It is thus necessary to assume that beside transfer from the substratum the developmental differences between HKE and SgE come into play and determine the usage patterns of V>N conversion in the contact varieties. It seems that the socio-institutional status of the English language in a region or country shapes the quantity of transfer from the substratum. (The range of transfer is also shaped by the degree of institutionalization of English, as the case studies in chapter 7 show.)

That the effect of the degree of institutionalization is so evident in the process of V>N conversion is due to the nature of conversion. It is a morphologically simple process that has been shown to be favored in early stages of language acquisition, i.e. in situations where target language proficiency is not so high as to allow for more complex word-formation processes such as affixation (cf. Pavesi 1998: 215). For Italian learners of English, Pavesi (ibid.: 226) finds that with an increase in proficiency comes a decrease in conversion. She traces the initial reliance on conversion back to its morphological simplicity on the one hand, but also its “economic motivation”: “a process whose meaning is predictable or readily recoverable from context does not need to be formally specified” (ibid.: 215). A tendency for linguistic economy has been noted for many ESL contexts as well (cf. Williams 1987: 169), examples

⁷The log odds as given in the ‘Estimate’ column are transformed into odds by applying the exponential function.

⁸A comparison between levels is not possible in such a logistic regression model. Individual levels can only be compared to the reference level, in this case British English.

include the well-known simplification of the English system of tense, mode and aspect or the omission of the copula *be*. This tendency for morphologically simpler structures seems to be intensified by the substratum and by the degree to which English is anchored in an ESL society. (Cf. section 9.4 for the problematic nature of the notion of *ESL variety*.)

The blocking constraint is universal

Another result which the model shows is that the frequency of the near-synonymous, deverbal noun is highly significant in determining the odds of V>N conversion. The higher the frequency of the deverbal noun is, the less likely V>N conversion becomes. This reflects the blocking constraint as demonstrated for USE in chapter 5. The constant tension between the creativity of the language user and the economy of language that blocks synonymous words is visible in the results. Nonetheless, it seems that these tensions play out differently in the varieties under scrutiny. In IndE, the blocking constraint appears to be considerably weaker than in all the other varieties. While blocking is equally strong in HKE, SgE, and BrE, IndE differs significantly from its parent variety. Adding the estimates ($-0.43 + 0.24 = -0.19$) one can see that the blocking constraint still applies to IndE but to a lesser degree. IndE is known for its liberal use of diverse word-formation processes and other means of vocabulary expansion such as borrowing (cf. Sailaja 2009; Sedlatschek 2009). In his work, Sedlatschek (2009: 145) finds “plenty of evidence [...] that users of IndE draw freely of the possibilities of borrowing, word formation and semantic change to expand their communicative possibilities and innovate their vocabulary”. Sailaja (2009: 40) notes that despite most IndE speakers’ orientation towards the British English norm in the morphosyntactic domain, the attitude towards innovative lexical items is more positive. In general, word formation in IndE seems comparatively unconstrained and speakers are used to dealing with high amounts of lexical variation, as studies on IndE word formation show (cf. e.g. Sedlatschek 2009).

Despite this amount of variation, it has been suggested that IndE displays a general preference for non-morphemic word-formation processes. Sailaja (2009: 82) finds that non-morphemic word-formation processes like “[a]bbreviations, clippings and acronyms [...] are plentiful in India” and that affixation, in contrast to compounding, is “certainly not as productive” (ibid.: 80). This could be an explanation why verb-to-noun conversion is fairly productive in IndE even though the typology of the main substratum would suggest otherwise.⁹ Furthermore, Mukherjee (2007: 175) and Biermeier (2008: 99) have also found a high

⁹According to Štekauer et al.’s (2012: 215) survey, Marathi and Telugu, two languages spoken in India, show conversion; however, only 7.0% and 7.2% of the Indian population gave these languages as their mother tongues in the most recent census (cf. Census of India n.d.[b]). It can therefore be hypothesized that even

productivity of V>N conversion in IndE. In Biermeier's (2008: 99) study, "Indian English has turned out to be particularly productive when it comes to coining new conversions", and Mukherjee (2007: 175) notes an increased usage of denominal conversions in IndE. The fact that IndE should favor an analytic word-formation process such as conversion dovetails with Kortmann and Szmrecsanyi's (2009: 280) finding that IndE scores very high on their transparency index, i.e. that IndE prefers transparent morphosyntactic markers over lexicalized forms.

The effect of verb frequency is only significant in HKE

A further result from the regression model is that the frequency of the verb itself, contrary to the initial hypothesis, does not seem to matter in BrE, IndE or SgE; it is not a significant predictor of verb-to-noun conversion. In contrast, verb frequency has a strong effect in HKE; it highly significantly predicts the odds of V>N conversion in HKE. The negative estimate indicates a negative relation. That means that the more frequent a verb is, the less likely verb-to-noun conversion becomes, and the less frequent a verb is, the more easily it is converted. This corresponds to what was hypothesized above. Frequent verbs do not show a greater autonomy in HKE. It is rather the low-frequency verbs that show higher odds of V>N conversion. Figure 6.1 illustrates how the frequency of the verb affects V>N conversion in HKE but not in the other varieties. The effect of verb frequency on V>N conversion is represented by lines for each variety. The line corresponding to HKE shows a steeper slope than the other lines. The ribbons along the lines represent the confidence intervals. Even though the confidence intervals are fairly wide, particularly for HKE, the result for BrE vs. HKE seems to be robust, as there is almost no overlap of the confidence intervals.

Discussion

This frequency effect can be interpreted as indicative of learner tendencies in HKE. In the high-frequency range, Hong Kong English shows less V>N conversion, which suggests that speakers repeat what they hear often and what is easily accessible. This, consequently, leads to a diminished success of verb-to-noun conversion in HKE for verbs of high frequency. Haselgren (1994) calls this tendency the *teddy bear principle* and Tschichold (2002: 133) describes it as follows: "learners clutch to what they feel is safe and familiar". It is firstly because of their increased frequency that frequent elements feel "familiar" and secondly because of their

though these substrates might influence the productivity of V>N conversion to a certain extent, this fact alone cannot conclusively explain the findings for IndE.

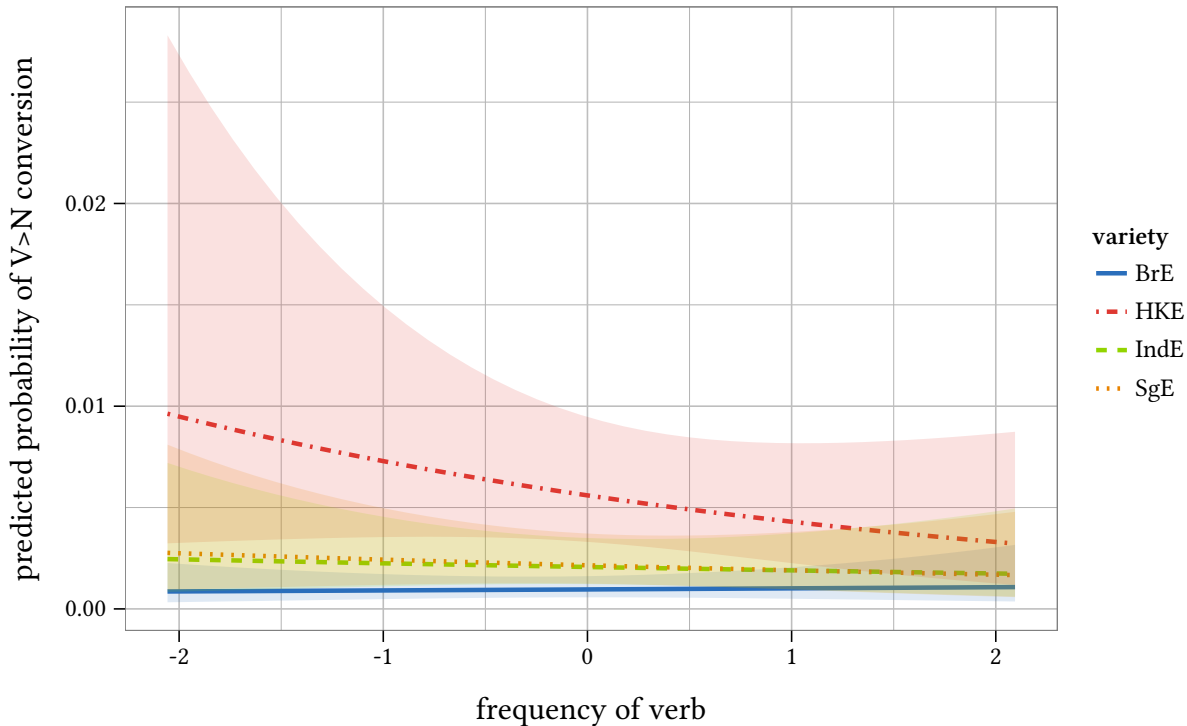


Figure 6.1: Highly significant frequency effect for verb in HKE

easy accessibility that they feel “safe”. The consequence of this is that high-frequency features of the target language “may be used even more often in ESL or EFL” (Biewer 2011: 15). Verbs of high frequency in HKE can thus be said to convert less easily, which is an indicator for the speakers’ reduced flexibility and creativity, typical of language learners.

In the low-frequency range, the opposite tendency can be found. Verbs of low frequency have higher odds of being converted than in other varieties. This once again tallies with findings on learner varieties. In forming a new word speakers seem to adhere to the easiest process that is available to them. To a speaker whose L1 is largely unfamiliar with derivational morphology, conversion is a suitable process. Wald (1993: 68) calls this the *shortest path principle*. Biewer (2011: 14) defines it as follows: “[I]f the rules of the target language allow for variation, one variant will be selected, and the selected variant will be the one that ‘correspond[s] most closely’ to the L1 feature”. This means that if English offers two possibilities of forming a new word—in this case conversion and derivation—it is more likely that speakers will opt for the process that they know from their L1. This is in line with Bao’s (2009: 346, 350) claim that only structures shared by the substratum and superstratum will re-surface in the grammar of the contact variety. Thus, an individual who regularly speaks a

Chinese dialect and whose proficiency in English is low is expected to favor conversion over other, morphological word-formation processes. Transfer in HKE is more extensive and also more likely for verbs of low frequency.

This tallies with Rohdenburg's (1996: 151) complexity principle, according to which more explicit formulations are preferred "in cognitively more complex environments", that is, e.g. with linguistic elements of low frequency, so as to reduce cognitive complexity and alleviate the processing burden (also cf. Hawkins 2004, cf. section 3.1.3). Furthermore, these findings are also in line with what Williams (1987: 178–179) calls the tendency for hyperclarity in learner language. Learners aim to establish isomorphism (one-to-one correspondence of word and meaning) and opt for maximum explicitness in their speech.

Summarizing these two tendencies, one can say that

[i]nput which can be processed more easily and effectively (because units and relationships can be identified more readily) [i.e. frequent forms] has a higher chance of being processed adequately, i.e. understood and then also replicated, and thus of becoming part of a newly emerging variety. (Schneider 2012: 64)

Hence, if a verb is very frequent, it is highly likely that there is a corresponding noun that is firmly established and that the speaker can use. Moreover, the speaker will be hesitant to use a form that is well entrenched as a verb as a different part-of-speech. (As has been pointed out for USE in the case of *connect*.) This verb is thus not likely to convert to a noun. If, however, on the other hand, a verb occurs with little frequency and a nominal form is not readily at hand, then the chances for this verb to be used in nominal contexts increase.

That this effect should manifest itself only in HKE and not in the other new varieties of English¹⁰ can be attributed to the developmental stage that HKE finds itself at:

It is immediately clear that SLA effects will be considerably stronger and more evident in the initial phases [of the Dynamic Model], marked by concurrent learning processes on the side of many indigenous people, as against the later phases when conventions have already become established in the speech community and SLA effects are thus more likely to be overridden by cultural conventions. (ibid.: 77)

IndE and SgE have both arrived at the stage of endonormative stabilization and their speakers' language proficiency can be assumed to be advanced enough to cope with these word-formation issues in a way that is fairly close to the native varieties.

The status of HKE as a variety in which substratum and institutionalization of English interact and consequently lead to a higher usage of conversion is also observed by Bunton

¹⁰Even though the estimates for SgE and IndE point in the same direction, the effect does not reach statistical significance.

(1991). In a study in which he compares mistakes made by international and Hong Kong learners of English he finds that “[t]here is a far greater tendency in Hong Kong to use a word of the wrong class than there is internationally”. As far as “the confusion of nouns and verbs” is concerned, “the tendency is [...] to use the noun: [...] **We must analysis the problem*” (ibid.: 19). This error is ascribed to the substratum, Chinese, in which verbs can be used as nouns without any overt morphological marking (cf. ibid.). However, he does not provide further comments on why this type of error is not to be observed with other English learners with Chinese as their native language. In the present study, the Dynamic Model is invoked to explain the difference between the two varieties with a Sinitic substratum.

The results of the first model can thus be said to confirm the hypotheses to a large extent. It has been shown that Asian varieties of English differ from the parent variety, BrE, in exhibiting higher odds of verb-to-noun conversion. Within the Asian varieties, HKE has the highest inclination for V>N conversion, followed by SgE and IndE. The effect of the substratum on the odds of conversion is moderated by the degree of institutionalization of English. This explains why the odds are significantly higher in SgE than in BrE but not as high as in HKE. The developmental stage further explains the higher odds of conversion in IndE, despite its mostly synthetic substrata.

Additionally, it has been shown that the blocking constraint is a tendency that is variety-independent and applies to English as a unified language. The more frequent a potential synonym is, the less likely it is that a new word is formed and established.

A global effect of the frequency of the verb which is to be converted on the odds of verb-to-noun conversion could not be determined. It is only in HKE that verb frequency becomes a significant predictor for the odds of conversion. In HKE, the more frequent a verb is, the lower the odds of conversion become.

The fact that the Asian varieties of English show such distinctive patterns as regards conversion calls into question notions such as *Asian Englishes* that describe and group varieties on purely geographical grounds. This has also been noted by Leimgruber (2013c: 5–6) for other varieties of English (also cf. section 9.4).

6.3.2 The globalized picture

In our globalized world, US American English has adopted a key position as regards the English language. As the dominant variety, it influences all other varieties around the globe (cf. section 1.1.3). The model presented in the previous section does not take US English into account. While parting from British English as the diachronic parent variety of HKE, SgE and IndE does have its validity, it also ignores a part of reality, namely the large influence of the

US English variety (cf. Mair 2013b: 261). In the following, another model is presented which adopts a globalized view. The input and variables in the model remain the same, except for the addition of US English data.¹¹ US English is consequently set as the reference level. The results are reported in table 6.4, additional coefficients are given in table D.2 in appendix D.2.

Table 6.4: Conversion in World Englishes

	Estimate	Std. Error	z value	p	
intercept					
(Intercept)	-6.58	0.17	-37.84	0.000	***
varieties					
GB	-0.15	0.07	-2.05	0.041	*
HK	1.51	0.09	15.95	0.000	***
IN	0.55	0.10	5.72	0.000	***
SG	0.64	0.12	5.34	0.000	***
frequency of deverbial noun					
frequencyDeverbal	-0.68	0.10	-6.56	0.000	***
frequency of verb					
frequencyVerb	0.37	0.14	2.61	0.009	**
variety : frequency of deverbial noun					
GB : frequencyDeverbal	0.23	0.06	3.81	0.000	***
HK : frequencyDeverbal	0.15	0.07	2.06	0.039	*
IN : frequencyDeverbal	0.45	0.08	5.75	0.000	***
SG : frequencyDeverbal	0.16	0.10	1.66	0.097	.
variety : frequency of verb					
GB : frequencyVerb	0.05	0.07	0.78	0.438	
HK : frequencyVerb	-0.26	0.08	-3.43	0.001	***
IN : frequencyVerb	-0.04	0.08	-0.56	0.575	
SG : frequencyVerb	-0.14	0.11	-1.33	0.185	

Trends persist

The results, displayed in table 6.4, are reassuring.¹² The main tendencies do not change when US English data are added to the regression model. Generally, BrE and USE do not differ

¹¹While it is highly unusual to input the same data set into two distinct models, in this case it is essential that the dataset only differ in whether it contains USE data or not. Otherwise, a comparison of effects for the ‘colonial’ in contrast to the ‘globalized’ setting is not possible.

¹²The prediction error of the logistic regression model (0.087%) is lower than that of the corresponding baseline model (0.18%).

much from one another. BrE is weakly significantly different in showing slightly lower odds of V>N conversion than USE. HKE, SgE, and IndE still differ highly significantly from the native varieties of English. HKE shows the highest numbers for V>N conversion, although slightly smaller than in the first model. Once again, SgE and IndE do not differ significantly from each other, as it seems. Furthermore, the blocking constraint can still be observed, yet it is considerably stronger.

Blocking constrains conversion to different degrees

The effect of the frequency of the deverbal noun is greatly increased ($B = -0.43$ for the ‘colonial’ model vs. $B = -0.68$ for the ‘globalized’ picture), which indicates that the blocking constraint is stronger in USE than in any other variety. This is confirmed when looking at the interaction between variety and the frequency of the deverbal noun. All varieties differ at least marginally significantly from USE and show a less strong effect of the blocking constraint. The variety that differs most significantly is IndE, as in the previous model.

Generally, the varieties seem to cluster into two categories: one in which the blocking constraint is highly effective and a second one in which it exercises a considerably lower influence on verb-to-noun conversion. Among the first group are USE, HKE, and SgE, whereas BrE and IndE fall into the second group. The blocking constraint is strongest for USE and the difference with SgE is only marginally significant. The log odds of V>N conversion increase significantly for HKE ($B = 0.15$) compared to USE. Nonetheless, the numbers for HKE are rather low, particularly when compared to the increase in log odds of BrE and IndE ($B = 0.23$ and $B = 0.45$ respectively). The fact that IndE is so similar to BrE is unexpected in light of the first model, where IndE shows a significantly higher resistance to the limitations imposed by the blocking constraint compared to BrE. In the first model, HKE and SgE do not differ significantly from BrE, whereas in the globalized model this does not seem to hold any more.

As it turns out, the difference in the strength of the blocking constraint disappears when CHOOSE is excluded from the model. Even though the prediction accuracy of the model is slightly lower, the blocking effect appears much more uniform if CHOOSE is excluded. All other effects remain the same, except for the fact that BrE differs less significantly from USE ($p < .05$). The fact that this verb should have such an impact on the model could be due to its status as the only verb that is not formed by suffixation but by apophony. The nominalization through ablaut might be processed differently than suffixation. This could explain the considerable effect that disregarding this verb in the analysis has on the logistic

regression model. Table D.3 in appendix D.3 gives the output of the regression model without CHOOSE.

Verb frequency is significant, even more for HKE

One major difference between the ‘colonial’ and the ‘globalized’ model is that in the model containing US data, the frequency of the verb itself becomes significant. For the ‘colonial’ model, the frequency of the verb only has an effect in HKE. The more frequent a verb is, the lower the odds of conversion to a noun in HKE become. In the globalized model, a different effect applies to all varieties: the more frequent a verb is, the more likely it is to convert to a noun. Nevertheless, this tendency is once again significantly different in HKE. There, the frequency effect of the verb is much weaker to the extent that it is almost non-existent compared to the other varieties. While the log odds remain positive for HKE in the globalized model ($0.37 + (-0.26) = 0.11$)—which means that the trend that more frequent verbs convert more easily to nouns is also observable in HKE—HKE can still be said to be distinctive in showing a considerably weaker frequency effect. It is the only variety that differs significantly from USE ($p < .001$). The other varieties show an overarching, non-variety-specific frequency effect.

For all varieties except HKE the general trend is that more frequent verbs convert more easily. Due to their higher frequency, these verbs might be more easily and hence more readily accessible also for nominal use than the corresponding derivation, which seems to lead to these forms converting frequently. This greater autonomy is somewhat unexpected, since it can generally be assumed that more frequent verbs are more strongly entrenched as belonging to the word class ‘verb’ and hence less likely to be used in non-verbal contexts (cf. hypothesis 3). Teddiman (2012) found that in an experimental setting with English native speakers ambiguous forms that can either be used as verbs or nouns (e.g. *work*) were classified as being a verb or a noun depending on the word class the form was predominantly used in. That is, ambiguous forms that are used as nouns more frequently had a higher chance of being classified as nouns rather than verbs and vice versa. In light of Teddiman’s (ibid.) findings, it is remarkable that all varieties investigated (except HKE) should favor verb-to-noun conversion for highly frequent verbs rather than low-frequency verbs.

That this effect shows up in this model could be due to collinearity in the data. The more frequent a verb is, i.e. the more often it is used in discourse, the higher are the chances that it also appears in its converted form. Nonetheless, since the number of converted forms contributes so little to the overall frequency of the lemma ($< 1\%$ for all verbs, except for CALCULATE and EXAMINE in HKE), collinearity is unlikely to explain the observed effect. It is

rather a higher discourse-pragmatic necessity that might help explain the effect. For highly necessary verbs, the pragmatic context should also require a corresponding noun. Since more frequent words are usually shorter than less frequent ones, converted verbs fit this pattern adequately: they express a nominal concept without the addition of any morphological material. Conversion produces comparatively short words which fit the needs of discourse.

Conversion as a last resort in USE

As far as US English is concerned, the model offers somewhat contradictory results. BrE differs slightly from USE, because it shows slightly lower odds of verb-to-noun conversion. (In BrE, the chances are 0.86 times those in USE.) The fact that USE should present a higher likelihood of V>N conversion is in line with for example Cannon (1985: 430). He states that “the process is producing large numbers of conversions” in USE, although most of these neologisms do not find their ways into dictionaries, which according to him is probably due to their nature as “slang items” (ibid.: 427). He says that “our functional shifts are noticeably more popular and less scientific than all the [other word-formation] categories”. This tallies with high numbers of converted forms in a web-derived corpus that supposedly contains language that exhibits characteristics of conceptual orality (cf. section 4.1.3). Nonetheless, the blocking constraint is extremely strong in US English, much more than in the other varieties. The picture that emerges is thus twofold: USE seems to favor V>N conversion, but only as a last resort, when a corresponding noun is extremely infrequent. While other varieties might accept the co-existence of two (almost) synonymous forms, one converted, one derived, the contrary is true for USE.

Blocking trumps verb frequency

The overall impression from the globalized model is that the two constraints on V>N conversion, the blocking effect and the verb frequency effect, are not of equal importance. While the verb frequency effect is already quite significant, with $p < .01$, the blocking effect is even more significant ($p < .001$) and also has a higher impact in terms of log odds ($B = 0.37$ for the verb frequency effect vs. $B = -0.68$ for the blocking effect). This tallies with the findings from the first model, where verb frequency yielded no effect but the blocking constraint was a highly significant predictor for the odds of verb-to-noun conversion.

6.3.3 Excursus: Refining the dataset

A closer look at the data revealed that some of the tokens, despite looking like nouns, could not easily be classified as such. In the following excursus, another model is presented from which these problematic tokens are excluded. The main contexts of these troublesome instances include contexts, first, where the supposed noun is part of a premodifier and, second, where the noun appears to be the result of a spelling mistake.

As for the first contexts, it is well-known that English is very generous when it comes to the structures of premodifiers. It even allows entire clauses in premodifying position. This tendency also seems to hold for the new varieties of English, as the following example illustrates.

(6.2) South Korean police have detained 26 confirmed, but has not yet been arrested a distribute indecent video man. (GloWbE-HK)

In this case, *distribute* occurs in premodifier position (premodifier underlined), but the entire premodifier exhibits clause-like characteristics. *Distribute* functions as a verb and *indecent video* as a direct object. Instead of opting for the relative clause (*a man who distributed indecent videos*), the author chooses to fit the relative clause into the premodifier. *Distribute* can consequently not be classified as a noun in this context.

Furthermore, in premodifier position, it is frequent that one finds instances of mention, i.e. words that are only mentioned but not really used in this context, as exemplified in the following.

(6.3) A traditional cause & effect diagram used for brainstorming future actions employed during the **Improve** phase of a DMAIC project. (GloWbE-IN)

In this example, *improve* occurs in a noun phrase and could be considered a premodifier to the head *phase* or even a modifier in an endocentric compound. Nonetheless, it is probably more accurate to assume that this is an instance of mention, rather than use. The capitalization of *improve* also hints in that direction. Any occurrences of potentially converted nouns in premodifier position, consequently, require a careful check of whether the token in question is an instance of mention or of use. In order to avoid any misrepresentation of data, in a second, more radical analysis all tokens with converted nouns in premodifying position were discarded.

Secondly, all nouns that may result from spelling mistakes were also excluded from the analysis. In some cases, such as the following, it seems plausible that the author simply misspelt a word instead of really making use of conversion as a means of creating a new word.

- (6.4) The words “taxes on the sale of goods” in Entry 48 mean taxes on a transaction the effect of Which is to transfer to a person for valuable **considers** tion, all the rights of an owner in the goods. (GloWbE-IN)

This example is an excerpt from a ruling by the Supreme Court of India. In this particular context, *considers* cannot be seen as an instance of conversion. Due to the legal nature of the text, one can conclude that the language must be of a rather high register and should therefore show no or only few non-standard features. In light of this, the appearance of the converted form is all the more striking. At second glance, the token has to be excluded because of a potential spelling mistake. Presumably, the author intended to type *consideration*. This becomes even more plausible when one considers that on a standard English keyboard <a> and <s> are adjacent keys. This token and others that instantiate such obvious spelling mistakes are excluded from the second analysis.

Choosing a more radical approach to data selection and sifting out all dubious instances also reduces the effect that might be produced by the extrapolation of frequencies. A misclassification of an individual token is multiplied, and thus has a much bigger impact, when the frequencies for the entire corpus are extrapolated on the basis of a sample of 1000 tokens. In order to avoid such mistakes as much as possible, the second dataset was created.¹³

After the re-codification of various nominal tokens to NAs, a total of 292 tokens remained (compared to 329 tokens before). The extrapolation then added up to a total of 1617 tokens. Figure 6.2 illustrates the changes in the data set. The blue part of figure 6.2 depicts the original dataset, the red part the modified dataset, with the x-axis indicating the counted numbers of nominal tokens and the y-axis presenting the values estimated by the logistic regression model. The closer the real and the estimated value, the closer the dot is to the black line. The red and blue lines are linear trends for the datasets. The grey ribbons represent the confidence intervals. The more the colored lines approximate the black line, the better the predictions are.¹⁴ As is apparent, from the first to the second dataset there is improvement, although rather subtle. This means that the coding of the first dataset was already fairly accurate and not much had to be re-codified in the creation of the second dataset.

¹³Note, however, that this procedure and the model that was subsequently calculated for the ‘colonial’ setting did not yield plausible results. This might be due to data scarcity that results from the deletion of various tokens. The fewer data points there are, the more susceptible the statistical model is to small changes in the data set. In the case of the ‘colonial’ model, this sensitivity to changes seems to be the cause for the implausible results that the statistical model yields. It is consequently not reported here.

¹⁴The model that is calculated on the basis of the second dataset shows a higher rate of correct predictions, i.e. a lower prediction error. While the value for the first model is 0.00087, the value for the second model with the restricted data set is 0.00078.

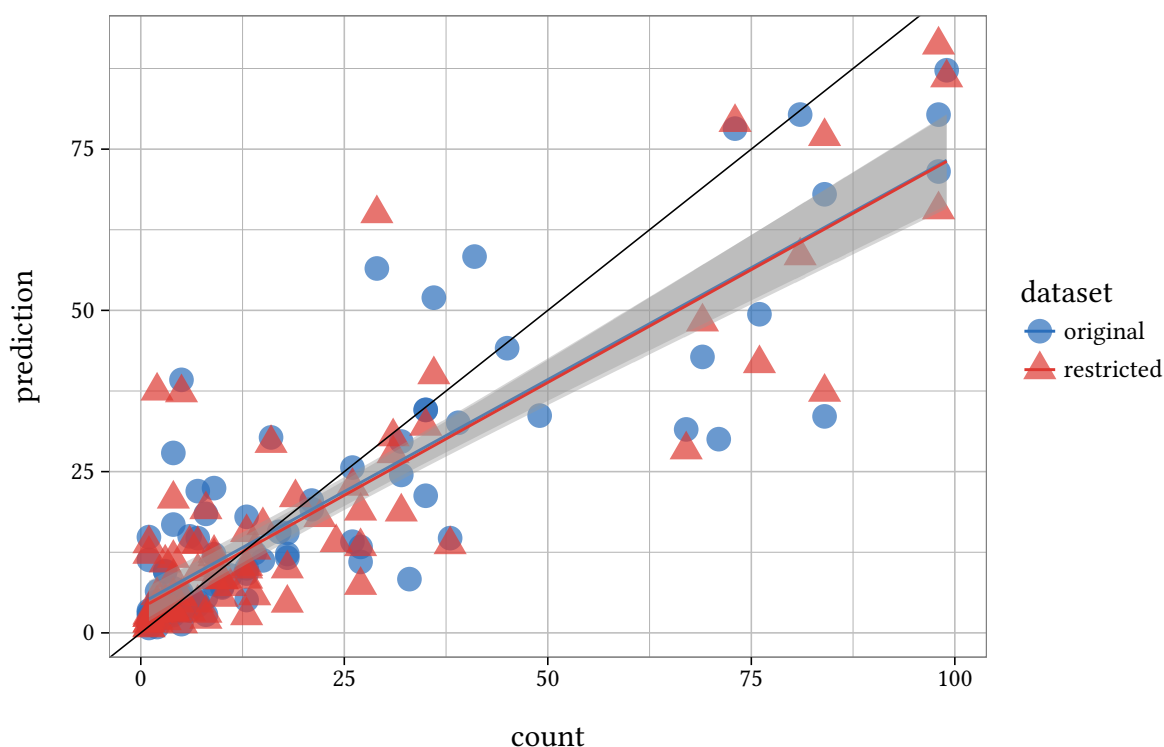


Figure 6.2: Datasets for the logistic regression. The datasets include all nominal tokens of all varieties.

The second dataset thus yields the results that are shown in table 6.5 when input into the same logistic regression model as above (US English is once again set as the reference). Additional coefficients are given in table D.5 in appendix D.4. The second model shows lower values for the AIC and BIC as the first model, indicating a better model fit.¹⁵ The overall impression is that the estimates are slightly higher in the second model, i.e. trends are slightly stronger. Nonetheless, this does not alter the general picture gained from the first model. The odds of verb-to-noun conversion are 4.68 times higher in HKE than in USE, 1.98 times higher in IndE, and 2.11 times higher in SgE. A critical difference between the models lies in that in the model with the restricted data set, BrE differs highly significantly from USE in showing lower odds of V>N conversion. Nevertheless, this effect does not, although highly significant, impact the model considerably. The odds of conversion are only marginally higher in USE than in BrE compared to the new varieties. The general picture thus remains stable in the sense that the native varieties show a much lower inclination towards V>N conversion than the new varieties. Furthermore, for the new varieties, the order of how high the odds of V>N

¹⁵AIC: 1429.1 (first global model) vs. 1345.0 (trimmed global model), BIC: 1470.8 (first global model) vs. 1386.7 (trimmed global model).

Table 6.5: Conversion in World Englishes (restricted dataset)

	Estimate	Std. Error	z value	p	
intercept (Intercept)	-6.80	0.18	-37.69	0.000	***
varieties					
GB	-0.36	0.09	-4.10	0.000	***
HK	1.54	0.11	14.59	0.000	***
IN	0.69	0.10	6.71	0.000	***
SG	0.75	0.12	6.04	0.000	***
frequency of deverbial noun frequencyDeverbal	-0.83	0.11	-7.76	0.000	***
frequency of verb frequencyVerb	0.49	0.14	3.43	0.001	***
variety : frequency of deverbial noun					
GB : frequencyDeverbal	0.48	0.07	6.66	0.000	***
HK : frequencyDeverbal	0.16	0.08	2.03	0.043	*
IN : frequencyDeverbal	0.52	0.08	6.21	0.000	***
SG : frequencyDeverbal	0.24	0.10	2.30	0.021	*
variety : frequency of verb					
GB : frequencyVerb	-0.11	0.08	-1.34	0.179	
HK : frequencyVerb	-0.33	0.08	-4.14	0.000	***
IN : frequencyVerb	-0.10	0.08	-1.27	0.205	
SG : frequencyVerb	-0.28	0.11	-2.55	0.011	*

conversion are remains the same, that is, HKE followed by SgE followed by IndE. Once again, IndE and SgE do not seem to differ significantly in the odds of conversion.

Another aspect of the model with the restricted data set is that there are slight differences to the previous model in the strength of the blocking effect in USE (stronger effect, $B_{model2} = -0.83$ compared to $B_{model1} = -0.68$) and also of the effect of verb frequency on the success of verb-to-noun conversion (effect stronger and more significant, $B_{model2} = 0.49$, $p < .001$, compared to $B_{model1} = 0.37$, $p < .01$). These differences between the models are only of a low magnitude, however.

For the restricted dataset, V>N conversion in USE seems to be even further constrained by blocking. All other varieties show higher odds of conversion, indicating a greater distance between those varieties and USE.

As far as the interaction of variety and verb frequency is concerned, the situation remains almost the same as above. HKE is still in a special position for not following the general trend of higher odds of V>N conversion for more frequent verbs. Despite the fact that the estimated log odds for HKE are positive, they differ highly significantly from those for USE, BrE, and IndE. (BrE and IndE do not differ significantly from USE.) What is new in the second model is that there is a tendency for SgE to also counteract this general trend. Nonetheless, it is not as strong and not as significant as for HKE. Since it is not as stable for SgE—it is neither observable in the first ‘globalized’ model nor in the ‘colonial’ model—it seems plausible that rather than constituting a robust feature of SgE this effect is due to the modified dataset.

6.4 Summary

In summary, what is apparent from the aforementioned regression models is that effects of frequency on the probability of verb-to-noun conversion hold in the same way across all models, although with minor variations in strength. Comparing the two ‘globalized’ models, it becomes evident that despite potential coding mistakes made in the first dataset the first model comes already very close to the second model, which is based on the ‘tidied’ dataset. If these models are then also compared to the model containing only BrE, HKE, SgE and IndE (the ‘colonial’ model), one can assume that those tendencies observable in all three models are the most robust ones.

Summarizing, what has to be pointed out is the unity of the English language despite its status as a globally used language. All tendencies which have been observed across varieties go in the same direction, whether it be the blocking constraint or the frequency effect for verbs. The log odds remain on the positive or the negative side of zero, effects are not reversed, they only increase or weaken.

The main global trend which is observed is the blocking constraint. In all varieties, the odds of verb-to-noun conversion are considerably lowered by an increasing frequency of occurrence of a potential derived synonym. The most striking local phenomenon is the verb frequency effect in HKE, where less frequent verbs show higher odds of verb-to-noun conversion compared to the other varieties, where less frequent verbs present considerably lower odds of conversion than more frequent verbs.

A further general trend seems to be that the blocking constraint ranks above the verb frequency effect. The verb frequency effect as a global tendency is only found in the global data and in the first of the globalized models it is not as significant as the blocking effect. Furthermore, across all models the verb frequency effect affects the log odds of verb-to-noun

conversion to a lower degree than the blocking constraint. This ranking of effects has also been demonstrated for the case of DISCONNECT versus CONNECT discussed in the previous chapter. Although CONNECT is by far the more frequent verb, it is not converted as often and as easily as DISCONNECT. This is due to the very reliable and very strong blocking effect that emanates from the deverbal noun CONNECTION.

7 A qualitative approach to conversion in Asian Englishes

After the quantitative data analysis has revealed a large-scale picture of conversion in World Englishes, it is indispensable to analyze select findings in more fine-grained case studies. The following three sections are dedicated to the three New Englishes, then various aspects common to all three varieties are discussed. The chapter concludes with a short summary which also briefly addresses the realization of V>N conversion in the native varieties.

7.1 Transfer from the substratum in Hong Kong English conversion

As the quantitative analysis has revealed, in HKE, the odds of conversion are higher than in the other two New Englishes, and it is therefore a process that is used comparatively frequently. On the one hand, this is the result of extensive transfer from the Chinese substratum, and on the other hand, the higher frequency of conversion is due to the learner effects present in this variety which come from the less institutionalized status of English in Hong Kong. The aim of this section is to shed light on conversion in HKE by qualitatively analyzing select examples. The excerpts stem from GloWbE and are complemented by data from ICE-HK in order to provide a more diverse overview of conversion in various text types, including the spoken medium.

7.1.1 Registers and formality

Considering that spoken and informal discourse is usually more progressive than writing and formal texts, it seems plausible that non-standard verb-to-noun conversion should mostly occur in the former contexts. However, the use of conversion in HKE is not only widespread in informal texts on the web but also in comparatively formal pieces of writing, such as the following excerpt from the webpage of the UNESCO.

- (7.1) Under the requirement and suggestions from UNESCO Beijing Office, the project of Experience of Universalizing the Nine-year Compulsory Education in Rural Areas in China was designed. It was expected that experiences in improving access to primary and junior secondary education can be summarized, and such experiences would provide useful lessons for other developing countries. # After the **approve** of the project of Experience of Universalizing the Nine-year Compulsory Education in Rural Areas in China, a few seminars were held by the office of College of Rural Education for Rural Development of Beijing Normal University or INRULED and an analysis framework and the timetable had been formulated.

Example 7.1 is an excerpt from an activity report on education in rural China by the UNESCO. Since the UNESCO is a well-known, globally operating organization, it can be assumed that a report from one of their project websites is comparatively formal in nature. The features of the text confirm this intuition. On the formal level, the text shows many abstract nouns (*requirement, suggestions, experiences*), which is characteristic of a learned style (cf. Biber 1989: 12). As far as the content is considered, the text is fairly technical in nature as it describes the goals of the project and the relevant procedures (*get approve, seminars held, analysis framework, timetable formulated*). Overall, apart from probably a tendency to use the passive voice comparatively frequently, the text displays few to no non-standard features, except for the converted form *approve*. What this example illustrates is that conversion in HKE can be found even in highly formal contexts such as the one presented here. It has to be kept in mind, however, that the conversion of *approve* is probably more readily carried out (and accepted) than other conversions since the suffix which is required to form the deverbal derivation /əɪ/ carries little weight compared to other derivational suffixes (such as /ment/ or /ʃ(ə)n/). *Approval* differs from the base verb in only two phonemes, one of which is a schwa.

The next example supports the finding that conversion occurs in comparatively formal registers by providing an excerpt from a business context.

- (7.2) After confirmation of your consultation, [our company will arrange business commissioner for measurement and **calculate** of volume] and [calculate the shipment time and the arrival time,] and [calculate the price depending on the details of your consignment items.]

Example 7.2 is from the webpage of a company called *Dragon Sea Shipping* that offers transportation of goods and moves to and from China (cf. Dragon Sea Shipping 2014). The excerpt is part of a step-by-step explanation of the moving procedure. The text shows a high density of deverbal nominal derivations such as *confirmation* and *shipment*. However, it also presents

one converted form that has not been suffixed. This instance of conversion could be due to the author of the text not being familiar with the noun *calculation*, a fact that is to be doubted considering the number of standard suffixations used in the text.

An alternative explanation is that this instance of conversion results from the structure of the sentence. The sentence consists of three coordinated clauses (*After confirmation...of volume, calculate the shipment time...arrival time, and calculate...items.*), two of which show a coordinated adverbial (*for measurement and calculate of volume*) and object (*the shipment time and the arrival time*), respectively, leading to four occurrences of the coordinating conjunction *and*. The author has probably mistaken the *and* coordinating the first and the second clause (underlined) for an *and* conjoining two elements of the head of a noun phrase and has therefore chosen to use the same form twice (*calculate*). The result is a converted noun and a verb in its base form.

Another option could be that this is an instance of spillover, but in the other direction, a ‘reverse spillover effect’. Spillover effects are usually encountered in reading tasks when speakers have read on but jump back, either with their gaze or mentally, to process what they have just read. This can cause a disproportionately long reading time for the subsequent element (cf. section 8.2). In this particular case, the opposite seems to happen. In production, speakers usually think ahead. In the present context, the author is probably already focussing on the next construction, a [V O] construction, which consequently leads to an influence of what follows, the [V O] construction, on the preceding construction, the [N of N] construction, yielding an instance of verb-to-noun conversion.

The next example again illustrates the use of conversion in a relatively formal context. The text recounts the “emergence of real trade unionism in Wal-Mart stores”.

(7.3) The Nanchang Bayi trade union was clandestinely set up on 14 August 2006. The chair, Gao Haitao, was elected by popular vote. Since then he had fought against Wal-Mart management over one issue after another. It is significant that he had studied law on his own while supporting himself by working at Wal-Mart part-time. In 2005 he passed a nation-wide **examine** in law and decided to stay on in Wal-Mart as a full-timer. His legal knowledge became his main weapon to fight against Wal-Mart.

Excerpt 7.3 stems from a webpage called *China Labor News Translations*, which offers “English translations of Chinese-language reports, commentaries and blogs on labor issues” (China Labor News Translations n.d.). The target group of the webpage are non-Chinese who are encouraged to “build a more nuanced understanding of how Chinese labor issues

are being conceptualised, understood and discussed in Chinese public forums”. The Chinese texts are translated by volunteers who are “former Chinese labor activists now residing outside China, the foreign media, or foreign scholars, NGOs, trade unions”. In summary, the website can be said to be directed towards (native) speakers of English and to serve an educational purpose (cf. China Labor News Translations n.d.). It is consequently not surprising that the excerpt should show an elaborate style marked by the use of infrequent lexical items (*clandestinely*), hypotactic sentences (*while supporting...*) and metaphors (*His legal knowledge became his main weapon to fight against Wal-Mart.*). Notwithstanding the formality of the text, the author-translator makes use of conversion in one instance (*examine*). Since this excerpt is a rendition of a text in the translator’s L1, it is highly likely that this instance of verb-to-noun conversion is due to direct transfer from the L1.¹ This non-standard use of conversion is all the more striking in light of the standard use of the present participle in the preceding sentence (*working*), which shows that the author-translator generally has a very advanced command of the English language.

The preceding examples have revealed that conversion is frequent in formal writing in HKE. Nevertheless, it also occurs in more informal contexts such as the one in example 7.4.

- (7.4) Their timings, characteristics and nature by itself, must be appropriate. When all these function favorable, ensure to do a background **examine** and avoid those with challenges.

Example 7.4 is from a text that offers advice on how to find a suitable room mate. The repeated use of the imperative (*ensure, avoid*) is a clear indicator of giving advice. Considering the interactive nature of advice, the communicative function of the text can be classified as largely appellative but also partly referential.

In short, as these examples demonstrate, conversion is not restricted to certain registers in HKE. It occurs in texts of all degrees of formality and with diverse communicative functions. Verb-to-noun conversion is thus a process that is not only pervasive as regards its frequency (cf. chapter 6) but also as regards the range of text types in which it is used.

7.1.2 Conversion in ICE-HK

The previous section has shown that conversion not only occurs in informal discourse but is also frequently used in comparatively formal settings in the written medium. In this section, examples from ICE-HK are drawn on to illustrate that the same holds true for the spoken

¹The omission of the article in *he had fought against ∅ Wal-Mart management* is a further indicator of L1 transfer.

medium.² The examples further suggest that conversion is a means of nominalization that is often exploited in situations where there is supposedly a comparatively heavy cognitive burden on the speaker. This is reflected in false starts, repetitions and the like. The following is a qualitative analysis of select examples in their discourse-pragmatic context.

- (7.5) A: And do you accept that the exchange rate of three dollars
 sixty five Hong Kong dollar with one Singaporean dollar was
 about five to eight in June nineteen eighty in uh June May
 June July nineteen eighty seven
 Z: I don't know
 → A: And there was refer to you you said earlier uh uh confirmed
 earlier on that uh there was not too much fluctuation in the
 (.) uh unit price invoice price either in nineteen eighty six
 and September nineteen eighty seven
 Z: Sales price

Example 7.5 is an excerpt from a legal cross-examination (ICE-HK S1B-062) in which speaker A is the cross-examiner and speaker Z the questioned person. In A's second turn, the converted noun *refer* appears in an utterance that apparently causes the speaker some difficulties. The utterance shows hesitation markers and also pauses as well as a false start (*you said earlier uh uh confirmed earlier*). These are not found in the preceding question in the first turn. It could be imagined that speaker A starts producing the second utterance while checking their notes for the statement that Z is supposed to have provided earlier. While A's attention is probably drawn to recapitulating Z's earlier statement, A reverts to conversion, a nominalization process that presumably imposes a lower cognitive burden on A than the process of derivation, since the former is transferred from A's native language. In addition, it could also be that speakers show a general preference for producing analytic variants in situations in which "processing demands are relatively high", as Kunter (2015: 35) found for native speakers' production of the comparative alternation.

- (7.6) A: And that that actually the choice they made will be subjected
 → to who's doing [the] the choose yes
 Z: [The choosing]
 Z: It subjects to availability and who to make [choice]
 A: [Yes]

The next example, 7.6, is an excerpt from a business transaction (ICE-HK S1B-079). In A's first turn, the speaker converts the verb *choose* to a noun after being interrupted by speaker

²For easier legibility, the original ICE transcripts are modified in such a way that they approximate Jefferson's transcription notation (cf. Jefferson 2004, appendix A).

Z. What is remarkable about this instance of conversion is that A still uses the non-standard form even after Z has prompted A with the standard form, a verbal noun in *-ing*, and after A has already shown that they know the standard form *choice*. It could be that the interruption by Z disturbs A to such an extent that A reverts back to a pattern from their L1, so as to alleviate the cognitive burden that results from being interrupted. This example furthermore suggests that A is not aware of their non-standard language use, seeing that A's turn finishes without any self-initiated repair. Even though Z's English proficiency and language awareness seem to go a little further than A's (apart from prompting A with *choosing*, Z also produces *choice*), Z does not initiate repair and inform A of their non-standard use of conversion. This could either indicate that Z is inferior or equal in status to A so that other-initiated repair would be interpreted as a serious threat to A's face that Z wants to avoid (cf. Brown and Levinson 1987: 61–68),³ or it could indicate that Z does not feel that the non-standard form *the choose* is worth initiating repair. If the latter were the case, it could only mean that Z is exposed to this kind of language use frequently.⁴

Excerpt 7.7, again from a dialogic business transaction (ICE-HK S1B-075), shows that speaker B's use of a converted form is embedded in a troublesome utterance.

³Considering that this is a business transaction, it seems highly likely that A and Z are two business partners who are on equal grounds as regards status and who furthermore behave deliberately politely so as to avoid face threats that could jeopardize the transaction. While this is a plausible scenario from a Western perspective, it is worth noting that the phenomenon of politeness can unfold differently in the Hong Kong context (cf. e.g. Schnurr and Chan 2009: 151–152).

⁴Alternative interpretations for the use of the converted form could be that (a) it is an instance of phonological priming by *who's*, or (b) that in the context of the conversation A interrupts him/herself and chooses not to finish their sentence seeing that the interlocutor Z has already understood what A wanted to say. Instead, A abbreviates their turn and supports Z's interpretation by uttering *yes*. I thank Ute Römer and Thomas Hoffmann (both p.c., 7 July 2015) for their thoughts on this excerpt.

- (7.7) B: You may in the you may have sometimes you may have to settle
for lower lower medical term for example I mean this <?> fat
</?> those are usually not the key elements [you consider] =
- A: [Uhm yeah]
= when you considering an employment okay
- B: If you can even if you can pick the choose okay [you] =
- A: [Yeah]
= won't okay you won't okay you won't take this as the fir-
uh first [priority] but then it it it just changes I mean =
- A: [Yeah]
= it just reduces your your your quality of life or your
quality of when you are when you si- when you [are] =
- A: [Right]
= contributing a sick okay is that you are going to a private
ward you may have to share a single ward with the SARS
[patient]
- A: [Oh]
- A: imprison no imprisonment well hospital yeah will be good one

The excerpt is about an employment option which speaker B does not consider “first priority”, so rather undesirable. It seems that the prospect of losing face due to overtly stating their opinion causes speaker B trouble. Indicators for this are, among others, numerous repetitions (*if you can, okay you won't, it, your*) and false starts (*fir- uh first*) as well as hesitation markers such as *uh*. In order to tone down their statement, speaker B repeatedly integrates mitigating discourse particles like *I mean* and *okay*. Even though B is not disturbed by A—through their continued backchanneling A actually shows fairly encouraging behavior— B is not able to fluently produce their utterance. In this context, B makes use of the converted noun *choose*. It is likely that the troublesome and potentially face-losing semantic content of the utterance requires B’s cognitive capacities to such an extent that B is unable to retrieve the noun corresponding to the verb *choose* and thus falls back on conversion as a nominalization process. This reading is even more plausible in light of the fact that the noun *choice* is not created by suffixation but by ablaut, an unproductive and very infrequent nominalization process (cf. e.g. Haselow 2011: 143).

As has previously been mentioned, verb-to-noun conversion also occurs in informal contexts. 7.8 is one example of these contexts.

- (7.8) B: But the requirement is like uhm you haven't studied in here
 for like te- ten years
 Z: Oh okay
 → B: Yah that's the major [require]
 Z: [Aw] you're not supposed to have
 studied in Hong [Kong]
 B: [Yes]

Example 7.8 stems from the private dialogue section of ICE (ICE-HK S1A-010). Here, similar to excerpt 7.6, B produces the converted form even though they know the derived form, as can be seen in B's first turn. However, because Z interrupts B, B is presumably presented with an increase in processing load and therefore uses the converted form, which is probably easier to process due to its similarity to the result of the corresponding L1 word-formation process.⁵

7.1.3 Syntactic contexts

Formal aspects

The examples from the preceding subsections indicate that conversion is used mostly in explicitly nominal contexts. Select examples of noun phrase constructions involving verb-to-noun conversion are repeated in table 7.1 for convenience.

Table 7.1: NP constructions with verb-to-noun converted forms in HKE

preposition	determinative	premodifier	head	postmodifier
after	the		approve	of the project
within	this		consider	
	a	nation-wide	examine	in law
	the	slow however steady	improve	
for	your		requires	
	their	fundamental	requires	
			choose ^a	

^a The entire clause is *I learn to make choose*.

Determiners that typically appear with the converted nouns are articles, possessive determiners and also demonstrative determiners. The converted nouns furthermore often occur in noun phrase constructions that form the prepositional complement of a prepositional phrase.

⁵It might also be that due to Z's interruption, B stops in mid-sentence. If that were the case, this is not an example of conversion but a plain anacoluthon.

Additionally, the converted forms are often used with premodifying adjective phrases or post-modifying prepositional phrases.

What all these examples show is that the use of converted nouns is not restricted to specific patterns, but that V>N conversion occurs in various different noun phrase patterns. Nonetheless, there seems to be a clear dispreference for bare or simple (i.e. consisting of determinative and head only) noun phrases. This could be due to the fact that bare or rather simple noun phrases do not mark conversion as explicitly as complex noun phrase constructions that involve pre- and postmodifiers. The more explicit the syntactic context is, the less the cognitive effort required for coercion (on the part of the hearer) becomes. If conversion were more costly to decode, it would most likely not be preferred over derivation. Considering that conversion by itself produces a word that is ambiguous and hence more difficult to decode, it is plausible that explicit contexts for conversion are preferred to keep up the decoding advantage of conversion over derivation (cf. section 3.1.3).

Yet, the question of why and how HKE speakers should be able to orient towards hearers and embed the converted form in an explicitly nominal context so as to ease the hearers' processing load in a situation in which the speakers themselves supposedly incur a high cognitive burden, remains unclear. A more detailed investigation of this phenomenon necessitates a larger corpus of spoken data.

Functional aspects

As regards the functions that noun phrase constructions with converted nouns as heads can fulfill in the clause, there are almost no restrictions. In HKE, converted nouns can appear in subject, object, and adverbial position. However, converted nouns as heads of phrases fulfilling the function of complements are not attested in the dataset. This clearly does not mean that this function is never assumed by a noun phrase with a converted noun. The absence of such instances could simply be due to the random sampling procedure with which the dataset was obtained.

Adverbials are often realized by prepositional phrases, which consist of a preposition and a prepositional complement, usually a noun phrase. Due to the nature of the prepositional phrase, the converted form is formally not the head of the prepositional phrase in which it occurs but the head of the noun phrase that forms the prepositional complement. Nevertheless, these occurrences are still considered instances of use as an adverbial. Table 7.2 gives select examples of each of the constituent types the dataset contains.

The qualitative analysis thus reveals that verb-to-noun conversion is a phenomenon of high pervasiveness in HKE. It occurs both in formal and informal registers, e.g. in UNESCO

Table 7.2: Clause constituents with verb-to-noun converted forms in HKE

subject	in such holy time, this choose is the good way to show true feeling each other moreover instantaneous well-being examine and also x-ray diagnosis are significant.
object	In 2005 he passed a nation-wide examine in law Sustaining poor develop [is the surest way to damage your body]
adverbial	After the approve of the project of [...] , a few seminars were held by the office of College of Rural Education for Rural Development of Beijing Normal University or INRULED Along with the continuously^a improve of product awareness , There will be more people share fast, efficient business experience in the future.

^a This is an instance of partial conversion. The form *improve* retains verbal qualities in that it is premodified by an adverb but also adopts nominal qualities by functioning as the head of a noun phrase that includes determinative and post-modification. This form, even though an instance of partial conversion, is still included because on a continuum between verb and noun it is considered to be on the 'nounier' side due to its predominantly nominal characteristics (as attributed by the syntactic context, i.e. article and postmodifying prepositional phrase).

reports and legal cross-examinations but also on webpages that offer advice in a more casual manner. That conversion should be employed in contexts so diverse as these is an indicator that HKE speakers might not be aware of the status of verb-to-noun conversion as a non-standard feature but rather see it as an acceptable nominalization process; a tendency that is most probably reinforced by the Chinese substratum.

As regards syntax, noun phrases containing converted nouns do not appear to be subject to any functional restrictions. These constructions can function as subjects, objects, and also as prepositional complements in adverbials. It seems that only formal aspects constrain conversion in HKE. The examples have shown that in HKE, verb-to-noun conversions are usually embedded in complex noun phrase constructions, so as to explicitly mark the converted form as a noun. Bare or simple noun phrases seem to be dispreferred, which is plausible considering that in these contexts V>N conversions are often ambiguous, and that reducing ambiguity through more explicit formulations reduces the processing cost associated with overriding the original word class.

7.2 Constraining transfer in nativization: Examples from Singapore English

After an analysis of conversion in HKE, this section aims to show that conversion as an effect of transfer from the substratum is constrained by an increased degree of indigenization, as is the case for SgE. In SgE, conversion is comparatively less frequent than in HKE despite the similar contact ecology of the two varieties. These differences are hypothesized to lie in the difference in socio-institutional status of English in the two speaker communities. In SgE, conversion is restricted to fairly informal contexts, as the examples in this section demonstrate.

7.2.1 Registers and formality

In SgE, not only the quantity but also the range of conversion is reduced compared to HKE. Conversion is less frequent in very formal contexts, but tends to occur in informal discourse in SgE. Some select excerpts shall help illustrate this claim in the following.

(7.9) You can say that I'm easily contented, no **deny** about that. I made a quick decision to be a Stay-At-Home-Mum five years ago and I went ahead to start an online business on my hobby two years back. Besides having the gut feeling and full support from my family, my positive attitude and optimism put me through those rollercoaster rides through these years. I do have my downtimes and bad hair days, but I've learn to pick myself up fast and keep moving forward.

The text in example 7.9 is part of a blog entry by Rachel Lim, a woman who describes her life as a mother on her personal website. Overall, her writing is very close to Standard English. Nonetheless, in the first sentence, she converts the verb *deny* to a noun. The determinative slot in the noun phrase is filled by a determiner and the postmodifier is introduced by the preposition *about*. Both characteristics, determiner and postmodifying preposition other than *of*, can be interpreted as being typical of a (deverbal) noun.

As is apparent in other instances of conversion as well, this particular example of verb-to-noun conversion could be the result of analogy, potentially modeling on the [*no N about*] construction. Usually, one would expect *deny* to appear in the form of a present participle in the idiom *no denying that*.⁶ However, due to the non-standard complementation patterns that SgE has been shown to exhibit (cf. e.g. Mukherjee and Gries 2009: 44), it is reasonable to

⁶In COCA, in 229 out of 329 times, the verb slot in the [*no V-ing that*] construction is filled by *deny*.

assume that the *-ing* form is non-preferable here. Consequently, the [*no N about*] construction is chosen, which helps avoid the use of a verb altogether. The analogy could further be fostered by the initial letter <d> that *deny* and *doubt*, the noun that fills the noun slot in this construction second most often,⁷ share.

Of further interest in Rachel Lim's post is her non-use of the past participle morpheme {-ED} in the last sentence. The use of the verb *learn* (underlined) in its infinitive form is an instance of simplification, a phenomenon that has repeatedly been attested for SgE (cf. e.g. Gut 2009; Terassa in preparation; eWAVE #132, Kortmann and Lunkenheimer 2013b). The simplification of *learn* in the past tense is consistent with the conversion of *deny* in the sense that both could be interpreted as part of a general tendency to avoid bound morphemes.

As far as the communicative function of the excerpt is concerned, Rachel Lim's text can be classified as fulfilling an emotive function. Her blog posts are personal in nature. She gives the reader insights into her daily life describing her experiences and emotions (e.g. "rollercoaster rides", "I do have my downtimes"), comparable to a diary. Her blog clearly falls within Grieve et al.'s (2010) personal diary blog type.

The next example, 7.10, is from a web shop which sells computer gadgets. The article gives advice on "What To Do With An Outdated Computer System".

(7.10) Hand-in-hand with cannibalizing is stripping out components and promoting them individually. This may be particularly helpful if you not too long ago installed a hardware **improve** to the now defunct system akin to a great DVD or CD reader/author, or a brand new onerous drive. Not only are the parts often easier to sell, they'll carry more money and ship much, a lot easier – and cheaper.

What is striking about this text is the comparatively elaborate vocabulary such as *defunct*, *akin*, and *onerous* and the use of complex syntactic constructions such as inversion (*Not only are the parts...*). Nonetheless, the author chooses to convert *improve* to a noun in the noun phrase *a hardware improve*. This form could be modeled on the analogous formation *software update*, a compound where the head also lacks a morphological marker that would indicate its status as a noun. Once again, on the formal side, the noun shows a determiner in the determinative slot and a postmodifying prepositional phrase introduced by a preposition other than *of*. The converted form further combines with a premodifier and thus fulfills all formal criteria for a noun.

⁷In COCA, the most frequent nouns that appear in this slot are *question* (1664 tokens), *doubt* (1567 tokens), and *mistake* (280 tokens).

The communicative function of this text is appellative and referential. The text both provides the reader with detailed descriptions of a computer but at the same times aims to offer advice on how to sell an old computer most efficiently.

In summary, conversion in SgE appears most frequently in texts with emotive and appellative functions, such as diary-like blog entries or recommendations on web forums that fulfill an interactive purpose. Even though some of these conversions are embedded in writing that shows little to no non-standard features, the context in which conversion is used remains largely informal as far as the communicative purposes of the texts are concerned. Conversion does not seem to occur in texts with a predominantly referential function. This contrasts with what is found for HKE, where conversion is also frequent in formal texts.

7.2.2 Conversion in ICE-SIN

What has been pointed out above is complemented in this subsection with data from ICE-SIN. The first finding is that conversion is much less frequent in ICE-SIN than in ICE-HK. Out of a handful of findings (among the 20 verbs under investigation), the following merits a detailed analysis, taking into account its pragmatic context.

- (7.11) B: No more extension
A: Then how now re-draw everything from scratch
B: Can submit old plans lah but it is like submitting the old drawings same old thing
→ B: Just that you know I mean except for this er FSB new require I mean FSB requirement lah
B: Know what I mean she's very terrible lah
B: Then her face will be involve and all that right
B: Then delay the construction

The excerpt in 7.11 stems from the private dialogue section of ICE-SIN (ICE-SIN S1A-051). In this text, speaker A and speaker B, probably students of architecture, are talking about plans for an object that have to be submitted within a certain time frame. In what comes before this excerpt they talk about the problems that one faces when one's plan does not fulfill certain requirements and does not get clearance. They further talk about how in order to revise a plan students can ask for an extension. Once a student is not granted an extension any more, the student has to "re-draw everything from scratch", as A notes. B then proposes the alternative that one could hand in old plans. In the subsequent and then following utterances B mentions the requirements again and how not fulfilling them comes

with certain problems (having to deal with a “very terrible” person/professor, potential delay of the construction).

Overall, this dialogue exhibits many features typical of spoken discourse. The text shows formally incomplete sentences (*no more extension, can submit old plans*), the Singlish discourse particle *lah*, hedges like *you know* and *I mean* and also hesitation markers like *er*. The troublesome turn in which B converts the verb *require* combines many of these markers. B starts the turn with two hedges, *you know I mean*, supposedly to now provide A with an extended explanation or comment on B’s suggestion to hand in old plans. Probably anticipating a following disagreement from A, B employs the strategy of hedging to mitigate the content of their utterance. B continues by referring to the *FSB requirement*. In doing so, B hesitates (as indicated by *er*) and then produces the converted form *require*. However, B quickly initiates a self-repair (cf. Schegloff et al. 1977: 364) with *I mean* and then produces the standard form *requirement*.

This excerpt illustrates that in ICE-SIN, similar to ICE-HK, conversion appears in situations that supposedly impose a cognitive burden on the speaker (in this case mitigating potential disagreement by A, building up the line of argumentation). Nonetheless, contrary to the examples from ICE-HK, the speaker immediately becomes aware of their use of a non-standard form and a self-initiated self-repair follows (contrary to excerpt 7.6, where Z provides A with a standard form but A still realizes the converted form). Even though this is only a single instance of language use that cannot be generalized, it still indicates that B’s command of English as regards V>N conversion is more native-like than that of many of the speakers in section 7.1.

7.2.3 Syntactic contexts

Formal aspects

The examples presented in this chapter show that in SgE explicit marking of novel converted nouns is preferred. In all excerpts, at least two slots of the noun phrase construction are filled. Table 7.3 illustrates this.

The picture for SgE is thus comparable to the one for HKE. Bare and simple noun phrases seem to be dispreferred, most likely due to the goal to encode conversion as explicitly as possible so as to facilitate processing, particularly for the hearer. However, the examples suggest a difference between the varieties as far as the prototypicality of the noun phrases is concerned. In SgE, the group of determiners also involves a negative determiner (*no*) and a complex determiner (*all the*). These are not found in the examples from the HKE corpora.

Table 7.3: NP constructions with verb-to-noun converted forms in SgE

preposition	determinative	premodifier	head	postmodifier
	no		deny	about that
to	a	physical	examine	
for		easier	refer	in the future
	my	own	expands	
	all the		requires	of her master

Consequently, conversion in SgE can be said to appear in comparatively less prototypical contexts and embedded in more complex constructions. Since less prototypical constructs are more difficult to decode, this could hint at the higher level of proficiency of SgE speakers compared to HKE speakers.

Functional aspects

The similarities between HKE and SgE are even greater for the functional aspects of noun phrase constructions with converted nouns. Comparable to what can be observed for HKE, noun phrases with converted nouns as heads can occur as various clause constituents in SgE. They are used as subjects, objects, and adverbials. Once again, there is no evidence of converted nouns being used as complements, which could be due to the chance nature of the data gathering process. Table 7.4 provides select examples of clauses containing constituents with converted forms.

Table 7.4: Clause constituents with verb-to-noun converted forms in SgE

subject	[Take time to understand] whenever your develop will be at its ideal for harvesting.
object	if you not too long ago installed a hardware improve to the now defunct system [that is evident in the manner that] the horse shifts direction and obeys all the requires of her master
adverbial	you can start bookmark those sites you like for easier refer in the future. paint your nail with your preferred nail improve

The pictures that emerge from the qualitative analysis of the syntactic contexts in which V>N conversion is used in HKE and SgE are thus similar. Conversion seems to be subjected to similar syntactic constraints. This is to be expected considering that HKE and SgE share a substratum. The main differences which can be identified between the varieties are thus the quantity of conversion (cf. chapter 6) and also the formality of contexts in which conversion is employed. In HKE, conversion is more pervasive than in SgE, whereas in SgE the greater degree of institutionalization presumably constrains conversion. The higher language proficiency which is expected to result from a higher degree of institutionalization is also visible in the conversational example in which conversion is immediately followed by self-initiated repair (example 7.11). Hence, the qualitative analysis confirms the results obtained by the quantitative analysis.

7.3 Liberal use of conversion in Indian English

The quantitative analysis has revealed that conversion is significantly more frequent in IndE than in native varieties of English. Unlike the situation in HKE and SgE, the high frequency of conversion in IndE cannot be attributed to transfer from the substratum to the same extent.⁸ It rather seems that novel verb-to-noun conversions are the result of a comparatively liberal and creative use of the word-formation process of conversion.

The examples in 7.12 and 7.13 illustrate this liberal use of conversion in IndE. In 7.12, the converted forms are used as bare nouns in subject position. In 7.13, the base of the converted nouns are phrasal verbs.

(7.12) Great experiences. When Baba himself is willing to give and help his devotees who can stop or reject. **Approve, disapprove** is in hands of Baba, humans are only his instruments and Baba himself is running the universe.

(7.13) One of the most simple methods you can take a step towards taking far better care of your teeth is to get a dental **verify up** with a dentist. I know, several individuals hate going to the dentist for fear of what will take place there, but I guarentee that you will be glad to get a **examine up** and a cleaning as soon as you have completed it. Look for a excellent dentist in your location in the newspaper, phonebook, by means of an online search, or by talking with friends. Just discover your self a dentist and make a visit. It is the greatest way to commence a lifetime of caring for your teeth.

⁸Cf. footnote 9 on page 144.

Example 7.12 is from a blog on “Devotees [sic] Experiences with Shirdi Sai Baba”, an Indian spiritual master. In describing his spiritual experiences, the author converts two verbs to nouns. What is remarkable about this particular instance of conversion is that it appears in subject position and as a bare noun without determinative, pre- or postmodifier. After having established that HKE and SgE disfavor bare converted forms, this example is all the more striking.⁹

Example 7.13 is from a website that provides the reader with advice on dental hygiene. The two instances of conversion in this excerpt are highly remarkable in that the base forms are phrasal verbs. Additionally, *verify up* and *examine up* are not established phrasal verbs; they are not listed in the OED and are also very rare in GloWbE, with *verify up* occurring a mere 10 times across all twenty varieties listed and *examine up* occurring 14 times. The forms in the example thus seem to be the result of two processes. The formation of the lexemes is most likely the product of analogy, that is, both phrasal verbs are modeled on the semantically similar phrasal verb *check up*. Probably also in analogy to the conversion process involved in yielding the noun *check-up*, the author of the text converts *verify up* and *examine up* from verbs to nouns. This combination of different processes shows the creative potential of IndE as regards word formation. What is more, *cleaning*, which is coordinated with *examine up*, is not formed by means of conversion, even though the coordination might invite this process, which would result in a structurally analogous formation. This illustrates that, seemingly, conversion is not used systematically in IndE.

7.3.1 Registers and formality

Similar to what is found for SgE, conversion occurs in comparatively informal contexts in IndE. The following example highlights this.

(7.14) 30-Aug-2013. Then why will it be in process? # I came us through company A and changed my job recently, my H1B transfer is in progress I did joined company with H1B transfer receipt. Now My Question is can I move to another job in this situation? # One more situation is could occur is, My current H1B (Which is in initial review state) goes into Rfe and I join my new Organization on based on the new Receipt then would it be a problem??? Or all the bridge H1B transfer has to get **approve** to apply /approval for a new H1B Transfer??

⁹It has to be noted, however, that this could also be an instance of *to*-deletion or of simplification (use of infinitive instead of present participle). Also cf. section 7.5.

Example 7.14 is from a webpage called *redbus2us.com* that offers guidance and advice on how to immigrate to the United States. The excerpt is from the comments section of this website. In it, the author is asking questions about the status of his visa and how it is affected by his working in a new job. The text exhibits a considerable number of non-standard features from a range of domains. There are instances of non-standard orthography (*anotther*, capitalization as in *Now My Question is...*) and punctuation (multiple question marks) as well as non-standard grammatical features. Among these are omitted prepositions and articles (*I came [to the] us [=U.S.], with [an] H1B transfer receipt*) and double past tense marking (*did joined*) as well as a novel verb-to-noun conversion (*approve*).

The comparatively high number of non-standard features in this short text points to the low language proficiency that the author seems to have. It is consequently not surprising to find a converted form as well. Another aspect which hints at a potential confusion the author is experiencing over the nominal form of *approve* is that almost immediately after producing the converted form the author is also able to retrieve the standard derived form *approval*. The fact that the author provides both forms and combines them with a slash, i.e. gives them as equal options, is also indicative of the author's insecurities in using the English language. The converted form further occurs in a frame that facilitates nominalization. Even though *get* is not among the semantically light verbs proposed by Dixon (2005: 459, 461), it is still unarguably a verb of low semantic weight (e.g. in the *GET-PASSIVE* construction, cf. Quirk et al. 1985: §3.66).¹⁰ In the present example, the use of the semantically light verb *get* makes the (pseudo-)nominalization of the verb possible, comparable to light-verb constructions.

7.3.2 Conversion in ICE-IND

As in ICE-SIN, conversion is extremely infrequent in ICE-IND in comparison to ICE-HK. The following excerpt from a broadcast discussion serves to illustrate the phenomenon of conversion in spoken IndE.

¹⁰This example could also be an instance of a non-standard realization of the *get*-passive.

- (7.15) A: [But] aging is not really a problem it's an issue (.)
 B: [Ahn]
 A: So wha- wha- what will be the impact on the economic health of the country (..) ?
 B: Basically Deepak-jee (.) uh once (.) you have a large population of people of sixty plus age group or age bracket (..) the dependency ratio starts increasing (.)
 → B: Many thereby should take the continue of life (.) from zero to hundred (..)
 B: There are more people at the younger age who depend on the people who earn (.)
 B: Similarly there is a larger group of people above the age of sixty (.) who're depending (.) on the earnings of the fewer people (.)

Example 7.15 is taken from a broadcast discussion (ICE-IND S1B-025) about the effects of demographic change and the increasing number of aging people in India. Speaker A, who is moderating the discussion, is asking speaker B, Dr Sharadchandra Gokhale, an expert on the topic (“the president of the International Federation on Ageing”) about the impact of aging on the economy. B then starts arguing his point and pauses frequently, possibly to build his line of argumentation. It is halfway through his argumentation that B uses a converted form. Presumably—similarly to what has also been shown for HKE and SgE—conversion is the product of the content of the utterance requiring more attention than this particular form. That is, explaining a fairly complicated matter in very well-structured sentences (his speech only shows pauses, no false starts or repetitions), imposes a heavy cognitive burden on the speaker. A higher attention on fluency leads B to subtract attention from exactly recalling words which then results in the converted form *continue*.

7.3.3 Syntactic contexts

Formal aspects

As far as the formal marking of converted verbs as nouns is concerned, IndE seems to be a case apart. As example 7.12 reveals, in contrast to HKE and SgE, conversion need not always be explicitly marked in IndE. The same tendency can be gleaned from example 7.16, from an article from “a website providing a press release distribution service” (Free-Press-Release Inc. 2013), in which a bare converted form is used.

- (7.16) Why More And More People Feel Like Driving Car To Somewhere Among different forms of transportation, cars have played an important role in our daily life bringing

us a lot of convenience, safety and entertainment. Due to this, more and more people have **choose**.

Nonetheless, the examples in table 7.5 show that, also in IndE, many instances of conversion are explicitly marked as nouns. The typical determiners that occur with these nouns are articles and possessive determiners, comparable to what is found for HKE. Both premodifying adjectives and postmodifying prepositional phrases are frequent. The phrase constructions in which conversions are typically embedded can thus be said to be as varied as in the other varieties, with the addition that IndE also allows for bare or simple noun phrases.

Table 7.5: NP constructions with verb-to-noun converted forms in IndE

preposition	determinative	premodifier	head	postmodifier
	the	scientific	examine	of dreams
	a	significant	improve	
on	the		require	from the client
	your	current	requires	
	his		requires	

Functional aspects

Contrary to the formal aspects of conversion, there are no differences between IndE and the Chinese-substratum varieties regarding the syntactic functions that noun phrases containing converted forms can realize. These noun phrases appear in subject, object, and adverbial position. As with the other varieties, there is no instance of their usage as complements, which is probably due to the random sampling procedure used for data collection. Table 7.6 exemplifies the various sentence constituents that can be acted out by noun phrases containing converted forms. As these examples show, the syntactic variability of conversion in IndE is similar to what is observed for the Chinese-substratum varieties. All Asian varieties permit noun phrases with converted nouns in a variety of syntactic slots, regardless of the contact languages.

The main difference between IndE and the other varieties thus seems to lie in the formal constraints that operate in HKE and SgE but not to such an extent in IndE. This is in line with other studies on IndE word formation that have shown that IndE is considerably more liberal when it comes to novel word formations (cf. e.g. Sailaja 2009: 75–84; Sedlatschek 2009: 145). Furthermore, it underlines the findings from the quantitative analysis, i.e. that the blocking

Table 7.6: Clause constituents with verb-to-noun converted forms in IndE

subject	The scientific examine of dreams is named Oneirology.
object	Due to this, more and more people have choose . <hr/> [I guarentee that] you will be glad to get a examine up and a cleaning as soon as you have completed it. <hr/> looking for schedules online has witnessed a significant improve in reputation .
adverbial	how will you count on points to occur in the choose ?

constraint is less effective in IndE, which also points towards greater flexibility in IndE word formation.

As far as register is concerned, conversion in IndE seems to be used mostly in comparatively informal contexts or in spoken discourse. This compares to what is found for SgE but contrasts with the findings for HKE where conversion is also frequently encountered in rather formal contexts. The fact that novel conversions generally do not appear in formal contexts indicates that while conversion is a frequent means of nominalization in IndE, it is clearly marked as being of a rather informal nature. Presumably, in more formal contexts the standard forms, i.e. derived nouns, are preferred. That speakers of IndE are able to distinguish between different degrees of formality of discourse and to employ different nominalization processes depending on the degree of formality hints at a high language awareness and proficiency, similar to SgE. These findings are in line with studies claiming that IndE shows an advanced degree of institutionalization (cf. e.g. Mukherjee 2007: 170).

7.4 Further observations

After providing detailed analyses for the individual varieties, some observations that concern all varieties shall be discussed in this section. These are conversions that have become lexicalized, conversions that are based on analogy, and conversions in light-verb frames.

7.4.1 Lexicalized formations

The first point worth considering is the number of lexicalized conversions. Many of the converted verbs conserve the original semantics of the verbal base (except for the reified

meaning). However, there are a few verb-to-noun conversions that seem to have adopted a lexicalized meaning. These cluster into two groups, those meanings that are similar to the original meanings and those that are new and have not yet entered the OED. Furthermore, there are various other converted forms that are used in contexts in which they acquire a non-standard meaning, yet, these are not systematic and must hence be assumed to be idiosyncratic usages.

Usage with related meaning

Two of the verbs among the twenty verbs studied display a more systematic and variety-independent tendency to be used in semantically different contexts. These are *examine* and *require*. In various varieties, *examine* as a noun is also used to mean ‘study’ or ‘check’ or ‘check-up’ (in a medical sense). The meaning of ‘study’ is shown in examples 7.17 to 7.19; examples 7.20 through 7.23 illustrate the meaning of ‘check’ or ‘check-up’.

- (7.17) The purpose of our **examine** was to check out the result of SYK inhibition on atherosclerosis. Our hypothesis [...] was based [...] (GloWbE-HK)
- (7.18) November 26th, 2012 admin # Prevalent Faults in Accounts Payable # [...] This **examine** lists the typical problems and errors relevant to accounts payable processing that happen to be present in most companies. # Common issues/errors # Data Entry Errors # Data entry mistakes can come about on any invoice field and account for some of the problems in accounts payable processing. [...] (GloWbE-SG)
- (7.19) A current research completed by the New York Instances says [...]. The **examine** takes this statistic and employs it [...] (GloWbE-IN)
- (7.20) We’ve an all in one **examine** that specializes in your luxurious adventure travel. (GloWbE-HK)
- (7.21) Ensure to do a background **examine** and avoid those with challenges. (GloWbE-HK)
- (7.22) Soon after this time duration once again these persons performed the all round health **examine ups**. (GloWbE-HK)
- (7.23) You will be glad to get a **examine up** and a cleaning as soon as you have completed it. (GloWbE-IN)

The converted verb *require* is sometimes used in contexts in which the noun *need* would be more appropriate, i.e. more frequently found. Examples 7.24 and 7.25 show these innovative uses of *require*.

- (7.24) In case you are seeking to get pleasure from your holiday as considerably as feasible, it is necessary that you simply [...] in order that you are going to have the ability to receive the types of items that you simply are in **require** of. (GloWbE-HK)
- (7.25) Gradually you will find no **require** for yoga classes and teachers. (GloWbE-IN)

Neologisms

Some converted forms have adopted meanings that are further away from the original verbal semantics than those presented above. The following examples illustrate this phenomenon.

- (7.26) Try out getting around 10 mins to accomplish **expands** when you are training. (GloWbE-GB)
- (7.27) Usually execute **expands** before getting into any workout or physical fitness process. (GloWbE-US)
- (7.28) [...] try out rubbing the muscles group close to that exact region. Do a few **expands** and use a warming cushion. (GloWbE-IN)
- (7.29) And third, as i mentioned before, it is quite possible to use a forge to defend your fast **expand** vs Terran at the moment, and it is a lot of fun to try out this new style, even if i feel like im more comfortable doing a Gateway opening vs Terran. (GloWbE-SG)
- (7.30) Heyy! i'm making some new **imagines** and was wondering if I could do a couple for you? (GloWbE-US)
- (7.31) There are two main challenges in shmups: The first is the '1CC' which is completing the game in one credit with no **continues**. The next is the high score. (GloWbE-GB)

Examples 7.26 to 7.28 show that *expands* seems to refer to the action of stretching muscles that is usually performed during physical exercise routines. In example 7.29, however, *expand* refers to the expansion of the user's territory in a computer game. The second meaning appears to be restricted to the computer game *StarCraft* and describes a special strategy for playing this game (cf. StarCraft Wiki n.d.). In example 7.30, *imagine* is used to refer to a piece of writing, as can be inferred from the blog from which this token is taken. A search

in the *Urban Dictionary* (Urban Dictionary LLC 1999), an open-source dictionary for slang words, reveals that *imagines* are “[a] type of fanfiction where the reader is included in the story as the protagonist” (kryzk 2014). Another creative use of language is given in example 7.31. Here, *continues* refers to gaming. According to the Wikipedia glossary of video game terms (Wikipedia contributors 2015-02-24), a *continue* is “[a] common term in video games for the option to continue the game after all of the player’s lives have been lost, rather than ending the game and restarting from the very beginning”.

Many of these instances of creative language use stem from the native varieties, which reveals that especially in the native varieties, conversion is a word-formation process that is mainly used in informal texts, in these examples particularly when referring to leisure activities such as sports, writing or gaming.

Some of the purportedly lexicalized converted forms turned out to be mistakes that resulted out of the confusion of the correct word with a converted form that is formally similar. 7.32 to 7.34 are examples of this. In 7.32 and 7.33, *imagines* is used instead of *images*, as is immediately clear from the context, which contains highly frequent collocates of *images* (*paint, conjure up*). Example 7.34 stems from a travelog and is about birdwatching. Clearly, *specifies* is mistakenly used for *species*.

(7.32) Painting allegorical **imagines** with words are the key to origin a dating site profile because people are looking for transcendence. (GloWbE-US)

(7.33) Amusingly the UCI’s press release says “this decision was made by the UCI together with all the implicated parties – in particular GCP...” which conjures up **imagines** of UCI staff consulting themselves. (GloWbE-GB)

(7.34) We had a Plane spotter minor problem with the camera today so unfortunately didn’t get a photo of all the different **specifies** but we hope we’ve sorted it now and the bird quiz will resume shortly!! (GloWbE-HK)

The fact that so few newly converted nouns adopt an independent lexical meaning indicates that verb-to-noun conversion is a word-formation process that is mostly used to facilitate the nominalization process and does not mainly serve to generate new lexical items. It is therefore not so much a word-formation but more of a word-form-formation process, so a rather grammatical process, to revert back to the traditional distinction between lexis and grammar for a moment. Predominantly in the native varieties is it the case that converted forms become lexicalized and that conversion is applied to describe entirely new concepts

such as pieces of fanfiction (*imagines*) or commands in video games (*continues*). Summarizing, the potential of conversion as a creative means in the lexical domain is higher in the native varieties, while in the new varieties, conversion is employed highly creatively in the grammatical domain or to refer to closely related concepts (e.g. *examine* to refer to *study*). Nonetheless, the latter instances are restricted to select lexemes. These findings mirror what has been observed for *disconnect* in chapter 5, where semantic shifts and the use of the lexeme with additional, metonymic meanings only start various years after the first occurrences of *disconnect* as a converted noun.

In Construction Grammar terms, the native varieties show a (still small but yet) greater potential for forming new atomic and substantive constructions that are characterized by an unpredictable meaning, whereas in the new varieties, the DEVERBAL CONVERTED NOUN constructions usually conserve the original meaning of the verb, except for the reconceptualization of these verbs as nouns. Conversion in the new varieties thus mostly serves the purpose of intra-paradigmatic normalization, i.e. regularizing the form of the noun to formally coincide with the verb.¹¹ In contrast, in select cases in the native varieties lexicalized nouns can emerge.

7.4.2 Analogical formations

Another aspect observable in all varieties is that analogy can serve as an instigator of conversion. Similar to example 7.9 for SgE, there are other instances of conversion that appear to have been formed on the basis of structurally similar constructions. One of them is *in require of*, which is presumably modeled on *in need of*, a construction that appears in all varieties. 7.35 exemplifies this analogy.

(7.35) If only part of your plan had been followed and monitored the small business would have already been lucrative and not in **require** of any outside **assist**. (GloWbE-SG)¹²

A further example is *examine up* as used in 7.36, mostly likely formed in analogy to *check-up*, which is frequently used in reference to a routine medical examination.

(7.36) I guarentee that you will be glad to get a **examine up** and a cleaning as soon as you have completed it. (GloWbE-IN)

¹¹This may be due to the circumstance that in new varieties, the morphologically less complex forms of the infinitive and the converted form are sometimes preferred over the more complex, inflected or derived forms in verb complementation and word formation (cf. section 7.5).

¹²It is noteworthy that the author of this text uses conversion twice in one sentence. Instead of *assistance*, the author prefers to use the converted form *assist*. This hints at a systematic use of conversion for nominalization purposes.

The non-standard orthography (*guarentee*) and the incorrect use of the determiner *a* suggest that this text is rather informal despite its comparatively serious topic (dental hygiene).

In some instances, conversion seems to result from priming (cf. Bock 1986). That is, a converted form is modeled on and therefore looks similar to another form that has been used in the context immediately preceding the converted form. Example 7.37 illustrates this phenomenon.

(7.37) Surely one of them could have warranted a mention – even a **refer** – on today’s front page. (GloWbE-US)

While *mention* is established as a noun, *refer* is not. Due to the close semantic proximity of the two lexical items, it can be assumed that the author of this text has again exploited the word-formation process that yields *mention* in order to nominalize *refer*.

7.4.3 Light-verb frames

As has been pointed out in section 2.1.3, it has been claimed for IndE that it favors light-verb constructions (LVCs, cf. Bernaisch 2015: 170–193; Hoffmann et al. 2011). In order to assess whether this holds true for all Asian varieties, all corpus samples from all varieties were analyzed with a view to their appearance in light-verb constructions. For this purpose, the concept of LVC was defined in a broader way than originally suggested in Dixon (2005: 462–467). Dixon (ibid.: 459) describes *give*, *have* and *take* as the core light verbs. He further mentions *make*, *do*, and *pay* (cf. ibid.: 461). In this analysis, all of these verbs were considered. Secondly, the LVC originally only allows the indefinite article (cf. ibid.: 459). For the purpose of this analysis, constructions that showed a zero-article or the definite article instead of the indefinite article were included as well. This was considered reasonable since new varieties of English often show usage patterns of the determiners that do not correspond to Standard English usage (cf. eWAVE, features #60 to #65, Kortmann and Lunkenheimer 2013b; Bernaisch 2015: 208 and Hoffmann et al. 2011: 267 for the zero-article in IndE). Light-verb constructions in which the converted form was either premodified or postmodified were excluded from the analysis, as such instances of conversion can clearly be recognized as nouns (cf. ibid.; Wierzbicka 1982: 755).

The results for the LVC with *have* are the following.

- (7.38) Among different forms of transportation, cars have played an important role in our daily life bringing us a lot of convenience, safety and entertainment. Due to this, more and more people have **choose**. (GloWbE-IN)
- (7.39) That's way SONY not update your Xplay. game or interface? have a **choose** ;p (GloWbE-GB)
- (7.40) He went to his doctor today to complain about some meds he was taking because he looses control of his bowels, the doctor told him he had a **choose** take the meds. or die. (GloWbE-US)
- (7.41) All your discount christian louboutin high heel sandals enable you to get able to be a part of this is what reasonably competitive current market place wish for direction ample this kind of have **consider** with the nation. (GloWbE-HK)¹³

Considering the number of converted forms that have been found in the corpus data (329 instances of verb-to-noun conversion), it is remarkable that only four should be embedded in this light-verb construction. It is even more noteworthy that this construction appears to admit one verb predominantly, *choose*, whose status as the only verb with a non-derivational nominal alternative (*choice* is formed by vowel gradation) has already been mentioned. It thus seems that the [*have* + (Det) + V] construction does not play a crucial part in verb-to-noun conversion but is rather restricted to select lexical items.

Furthermore, the semantic criteria for these lexical items only partly overlap with Wierzbicka's (ibid.) criteria. According to Wierzbicka (ibid.: 759), "[t]he *have a V* construction is agentive, experiencer-oriented, antidurative, atelic, and reiterative" (also cf. Dixon 2005: 469–470). The first two criteria are fulfilled by the *have (a) choose* construct. It is agentive, that is, a person carries out the act of choosing. Secondly, it is experiencer-oriented in that the effects of the choice immediately impact on the person choosing. Notwithstanding these characteristics, the *have (a) choose* construct does not fulfill the other criteria. By "antidurative", Wierzbicka (1982: 757) means that the action described by the converted form "cannot be momentary: it must go on for some time", but not for an extended period of time. The verb *choose*, however, is not atelic and the action of choosing usually does not "go on for some time". What is more, it is to be doubted whether the *have (a) choose* construct fulfills the criterion of being reiterative. In example 7.40, the act of *choosing* cannot be considered

¹³This text seems to be a (possibly computer-generated) advertisement for shoes rather than an actual blog entry. It is therefore not analyzed.

reiterative due to the fact that the choice is between a terminal illness and medication, which is supposedly a decision that can only be taken once.

Consequently, the [*have* + (Det) + V] construction cannot be thought of as facilitating verb-to-noun conversion to a considerable extent. It rather seems that the *have (a) choose* construct has become an independent construction that has taken on characteristics usually not attributed to the [*have* + (Det) + V] construction. Seeing [*have (a) choose*] as a construction in its own right would also explain why *choose* occurs in the LVC with a higher token frequency than any other verb.

As regards other light verbs, the following constructs could be extracted from the corpus data.

(7.42) Add onto that some fully body works far better for us due to the fact is quite content in the market today then I do not cause you the inspiration to take into **consider**.
(GloWbE-HK)

(7.43) From this, I learn to make **choose**. (GloWbE-HK)

(7.44) That could be better, well every young boy always take **refer** from this site.
(GloWbE-US)

The results reveal that other light verbs are encountered even more rarely. Due to the fact that only the last of these examples can be traced back to a website that still exists (*Men's Health* magazine), the other examples should probably be analyzed with considerable caution. While these examples provide too little evidence to allow for any kind of generalization, it should nevertheless be noted that the fact that two out of three stem from the HK section of GloWbE could be taken as an indicator that HKE is more flexible than the other varieties in accommodating such light-verb constructs.

Accordingly, the systematic analysis of light-verb constructions reveals that, contrary to what was hypothesized, verb-to-noun conversion is comparatively infrequent in LVC contexts. Light-verb frames do not seem to facilitate verb-to-noun conversion, at least not in the texts in GloWbE. This could be due to the fact that the web-based texts are comparatively informal and thus use other means of marking conversion, e.g. by embedding them in noun phrases. In some new varieties, the LVC appears to have lost its marking as a colloquial construction (cf. Hoffmann et al. 2011: 271), which could explain its non-occurrence in the informal web registers.

A further explanation as to why LVCs are not as frequent in the web data as initially assumed comes from a neurolinguistic study. Wittenberg et al. (2014) found that light-verb

constructions are more difficult to process than the corresponding non-light constructions, regardless of the frequency of the LVCs or the verbs used in them. Participants consistently showed an N400 negativity effect that “reflects an extended process of integrating the incoming word” with the preceding context (ibid.: 40). They conclude that this must be due to the higher complexity of LVCs compared to non-light constructions. LVCs, they claim, show “argument sharing”, that is, the agent of the verb is at the same time the agent of the converted form (e.g. of *give* and *kiss* in *give a kiss*), and the patient is the patient of both actions expressed by the verb and the converted form (e.g. the recipient of *give* and also the patient of *kiss*, cf. ibid.: 31). As LVCs require the hearer to decode and resolve the argument sharing, these constructions come at a higher processing cost and hence do not constitute a processing advantage over non-light constructions. This could be a reason why LVCs are not preferred in these contexts in new varieties of English.

7.5 Locating conversion on the lexis-syntax continuum

Since there is considerable resemblance of the findings for all new varieties as far as the lexis-grammar continuum is concerned, this section summarizes the results for all the Asian varieties. In HKE, SgE, and IndE, there is a tendency to realize verbal inflection in non-standard ways as the following examples demonstrate.

- (7.45) The inquisition conclusion may be different when **choose** the different brand.
(GloWbE-HK)
- (7.46) This flat rate allowance is provided because uh uh the Inland Revenue want to reduce the time in **examine** those numerous frames on uh small items on small items.
(ICE-HK, S1B-015)
- (7.47) Hear this news, the happiest person was not domestic club boss to not be belonged to, which boss does not hope your writing saves cost of **choose** and **employ** persons?
(GloWbE-SG)
- (7.48) The main purpose of social bookmarking is for people to book mark their favourite and get their bookmark websites store at the bookmarking site. If you were to use social bookmarking site to store your own favourite websites, choose a site you prefer and register an account. Take some time to learn about the site and you can start **bookmark** those sites you like for easier **refer** in the future. (GloWbE-SG)

- (7.49) Tools such as Radian 6 can listen and digest social sentiment by **understand** things like how many liked a product or didn't like it. (GloWbE-IN)
- (7.50) If business is carried out without planning it will not bring desired results but will definitely bring undesirable results. Planning decides the future course of action of business by **consider** all factors, which can be influence the action. (ICE-IND, W1A-016)

From examples 7.45 to 7.50 it is apparent that speakers of all new varieties, HKE, SgE and IndE, have difficulties in producing the present participle form of verbs. This can either be attributed to extensive transfer from the analytic substratum (for HKE and SgE) or to a general simplification tendency common to many L2 varieties of English.

The first explanation—that verb morphology is reduced in those varieties with an analytic substratum—is illustrated in 7.47. This example is an excerpt from a web page on hiring players and other employees for sports clubs. The verbs *choose* and *employ* complement the noun phrase head *cost*. While in standard English one would expect a gerund in this slot, the speaker of SgE uses the infinitive instead.¹⁴ For speakers of SgE and HKE the use of this verb form is probably attributable to transfer from the Chinese substratum, a language that does not show verbal inflection (Bao Zhiming p.c., 8 July 2014 for Mandarin, cf. Chan 2010: 305 for Cantonese). Consequently, the standard-like use of the gerund or present participle is likely to pose a major challenge for speakers of English whose native language is Chinese. In a study with intermediate and advanced Hong Kong learners of English, Chan (ibid.: 305, 308) found that even advanced learners (university students) still exhibited considerable rates of non-standard verb forms as well as of “word class confusion”, which she concludes are most likely due to the influence of the analytic Chinese substratum.

As far as IndE is concerned, transfer from the substratum can largely be ruled out as an explanation for the non-production of the present participle. The second explanation, simplification, must thus be explored. The simplification of verb morphology is a phenomenon which is common to all new varieties regardless of their contact ecologies. Szmrecsanyi (2009: 328–329), for example, found that the L2 varieties studied obtained a lower overall value on the syntheticity index, i.e. made less use of bound grammatical morphemes, than the native varieties. Furthermore, IndE scored higher on the analyticity index than other L2 varieties, which means that despite the fact that the substratum is unlikely to foster the simplification

¹⁴The fact that the object (*persons*) immediately follows the verbs (*choose and employ*) and is not embedded in a prepositional phrase is a clear indicator that these are verbal uses of the infinitive and that the verbs are not used as deverbal nouns here. That is, *choose* and *employ* are not instances of conversion, even though they might have been analyzed as such by the tagging software used for the preceding quantitative analysis.

of verbal forms in IndE, IndE in itself shows a preference for analytic grammatical markers compared to other new and native varieties. This can help explain why the present participle is not always realized in IndE where it would be in standard varieties.

What is further evident from these examples is that, functionally, the continuum between syntax (i.e. inflection) and word formation (i.e. verb-to-noun conversion) does exist in all varieties but that, formally, it is realized very differently from the way it is realized in the native varieties of English. Example 7.48 is drawn on as an illustration.

Example 7.48 stems from the comments section below an article that describes and explains social bookmarking. The author of the article, Moon Loh, replies to a reader's comment. The excerpt presents two instances of nominalization. The first is *bookmark*. The position that *bookmark* occupies in the clause is typical of a present participle. In the main clause, it assumes the function of the object dependent on the main verb *start* (as part of the subordinated clause *bookmark those sites ... the future*). Within the non-finite subordinate clause, it clearly is the main verb upon which the object of the subordinate clause (*those sites you like*) depends. Its verbal characteristics are further underscored by the non-occurrence of a determinative or postmodifier. What is untypical is that the verb is not marked as a participle by the inflectional morpheme {-ING}. Nonetheless, because of the verbal characteristics of the form, it cannot be interpreted as an instance of a deverbal noun. *Bookmark* can thus be seen as an example of non-standard verb complementation.

The second instance of nominalization in this example is *refer*. It shows clearly nominal formal characteristics. It is the head of a noun phrase in which both the premodifier and the postmodifier slot are filled. The premodifier is an adjective (*easier*), that is, pertains to a word class that can modify nouns but not verbs. The postmodifier is a prepositional phrase that is not introduced by the preposition *of* (*in the future*). *Refer* consequently fulfills the criteria for deverbal nouns.

Thus, even though the new varieties exhibit the same functional spectrum, formally, these cases are either realized as infinitives (e.g. *bookmark* in example 7.48) or as deverbal nouns (e.g. *refer* in example 7.48). In native varieties, the gerund or derived nouns are expected to be preferred in these slots. Hence, the formal side of the continuum can be said to be less elaborate in the new varieties than in the native varieties. That SgE and HKE should present a verb morphology that is less complex than the standard is probably to a large extent due to the Chinese substratum, which does not distinguish between verbs and deverbal nouns. Furthermore, for IndE as well as the other varieties, there is a general tendency to simplify complex constructions, which seems to apply to the present participle and gerund too.

As it appears, complex constructions such as the PRESENT PARTICIPLE construction pose problems for speakers of new varieties (even more so for speakers with an analytic L1) and are therefore avoided. This leads to the omission of the {-ING} suffix as shown above on the grammatical end of the continuum and to a preference of conversion over derivation on the lexical end of the continuum (compared to the native varieties). Even though the result looks identical, the forms belong to different constructions. While the use of a verbal infinitive instead of a present participle is an instance of non-standard verb complementation, converting a verb to a noun generates a new construction since the converted noun (usually) adopts all the features characteristic of that part-of-speech.

Due to the fact that more complex and more schematic constructions such as the PRESENT PARTICIPLE construction [V *-ing*] or derived nouns, e.g. the [V *-tion*] construction, are more costly to process, these constructions are underused or even avoided in Asian Englishes by either omitting suffixes or converting forms, embedding them in contexts that clearly require a different part-of-speech. The converted forms look like atomic and substantive constructions and can therefore be expected to be processed with greater ease. Table 7.7 summarizes the findings on the lexis-grammar continuum contrasting Asian varieties with native varieties.

Table 7.7: The lexis-grammar interface in Asian Englishes. Shaded forms are underused compared to BrE and USE.

grammar core syntax		←————— complementation —————→					lexis core word formation	
traditional grammar:	infinitive	present partici- ple	gerund	verbal noun <i>-ing</i>	in	deverbal noun by conversion	derivation by suffixation	
construc- tion type:	atomic and substantive	complex schematic	but bound,	partly	atomic and substantive	complex but bound, partly schematic		

7.6 Cross-varietal differences in register

The Asian varieties analyzed in this study have not only been found to differ as regards the frequency of occurrence of verb-to-noun conversion. The qualitative analysis of this phenomenon has revealed that across varieties conversion occurs in different contexts. In all varieties, conversion is used in informal contexts and spoken discourse. However, in

HKE, conversion is furthermore comparatively frequent in much more formal contexts, as the example of a UNESCO report (cf. example 7.1) illustrates.

For HKE, it thus seems that speakers frequently fall back on a process from their L1, transferring structures from the substratum regardless of the formality of the context. In a study on Hong Kong students' oral and written production, Chui (2010: i) found that Hong Kong students were unable to adequately employ linguistic resources available to them. Overall, the students' written texts exhibited more features characteristic of informal, spoken speech than native speakers' written texts. This could explain why verb-to-noun conversion, a phenomenon typical of informal contexts in standard and advanced varieties of English (as it seems), is still used in formal contexts in HKE. Lower overall language proficiency and language awareness result from the lower degree of institutionalization of English in HKE. This results in an increase in transfer from the substratum and in an increased usage of non-complex noun constructions that are easy to encode and decode.

In SgE, on the other hand, institutionalization has advanced to such a degree that speakers mostly restrict their usage of verb-to-noun conversion to informal contexts such as the diary-like weblog entry in example 7.9. Similarly, in IndE, conversion is largely limited to informal contexts. This is in line with what has been reported on the use of conversion in native varieties of English (cf. Cannon 1985: 427). It further tallies with the observation made in the study of *disconnect*, namely that the recently converted form is preferred in more informal registers (cf. chapter 5).

7.7 Summary

The close scrutiny of conversion in Asian varieties of English has revealed that the differences between Asian varieties not only manifest themselves at the quantitative level (cf. chapter 6) but are also present at the qualitative level. While verb-to-noun conversion does exist in all three varieties, the productivity of the process in the individual varieties is subject to variety-specific constraints predominantly affecting form and register.

It has been pointed out that HKE occupies a special status in the group of Asian varieties in that it allows verb-to-noun conversion not only in informal but also in very formal registers. In IndE and SgE, on the other hand, the process is restricted to mainly informal contexts. This has been attributed to the lower degree of institutionalization of HKE compared to IndE and SgE, seeing that the more advanced varieties (SgE, IndE) show profiles similar to the native varieties. Summarizing the findings for British and American English, it can generally be noted that even in the plethora of text types which the internet encompasses conversion

is an elusive phenomenon in native varieties of English. It is found in contributions to forums and discussions where users interact responding to one another. The content of the messages is often personal; the texts have an emotive or appellative function. Generally, the verb-to-noun conversions in BrE and USE are often spontaneous ad-hoc formations, used in contexts that show little planning, as is evident from the number of orthographical as well as grammatical mistakes.¹⁵

As far as formal constraints are considered, HKE and SgE present a clear preference for the embedding of verb-to-noun conversions in explicitly nominal contexts, e.g. preceded by a determiner and/or a premodifying adjective. This parallel could result from the shared substrate language of HKE and SgE, Chinese. In contrast, in IndE, the phenomenon can also be embedded in bare noun phrases, which hints at a higher tolerance of conversion in IndE. This is in line with the ‘liberal use’ of various word-formation processes that has been mentioned in the literature (cf. Callies 2015; Sailaja 2009: 75–84; Sedlatschek 2009: 145) and also with the reduced strength of the blocking constraint that the quantitative analysis has revealed (cf. section 6.3).

The occurrence of the DEVERBAL CONVERTED NOUN construction in largely explicitly nominal contexts, at least in HKE and SgE, consequently begs the question whether the grain size of the construction is actually larger than what was originally assumed in chapter 1.1.2. The results presented in this chapter might suggest the existence of two constructions, first, the original DEVERBAL CONVERTED NOUN construction $[V]_N$, and second, the NOUN PHRASE construction with a deverbally converted noun as the head $[\text{Det}_{\text{dtm}} X_{\text{premod}} V_N Y_{\text{postmod}}]$. Due to the fact that a qualitative analysis necessarily has to focus on a handful of examples and exclude others from the analysis, this question cannot be answered conclusively on the basis of the data provided here. As it seems, all varieties allow bare nouns, that is, the $[V]_N$ construction, but to different degrees, with IndE taking the lead. In addition, many of the converted forms occur in the $[\text{Det}_{\text{dtm}} X_{\text{premod}} V_N Y_{\text{postmod}}]$ construction, which presents a broad variety of pre- and postmodifiers as well as determiners. Thus, considering that there is no unique pattern for any of the Asian varieties (e.g. non-occurrence of the bare form, occur-

¹⁵A further example of V>N conversion is given in 7.51, which is also a post in a comments section. It is from a blog that discusses a widely used cell phone from a globally operating company.

(7.51) # pegel # calm down man!!, if you want ICS come to your X Play don't wait for SONY update, just Install costume room with ICS., and say goodbye for game, that's way SONY not update your Xplay. game or interface? have a **choose**; p # (GloWbE-GB)

Non-standard, inconsistent orthography and punctuation indicate the informality of the text. An emoticon at the end and also the fact that the writer addresses another user as “man” further point in that direction. The paratactic sentence structure contributes to the impression that conversion is fostered by the informal nature of the text.

rence mostly/exclusively with premodifier) it can be assumed that the DEVERBAL CONVERTED NOUN construction indeed takes the form of $[V]_N$, even though in the majority of contexts this construction is greatly attracted to the NOUN PHRASE construction, into which it is frequently embedded.

In order to further disentangle substrate transfer effects and learner effects and to ascertain how conversion is processed by speakers of different varieties, an experiment was conducted. The experiment design and the results are described and interpreted in the next chapter.

8 Experimental validation of corpus results

Both the quantitative and the qualitative corpus analysis have revealed considerable differences between varieties of English in the frequency and also in the contexts of use of verb-to-noun conversion. Following Schmid's (2000: 39) *From-Corpus-to-Cognition Principle*, it is assumed that "frequency in text instantiates entrenchment in the cognitive system". By entrenchment, Schmid (2007: 119) means "the degree to which the formation and activation of a cognitive unit is routinized and automated". Nevertheless, the link between frequency and entrenchment is not a direct one. Schmid (2010: 116–117) explains their relation as follows.

[W]hat frequency counts in corpora reflect more or less directly are degrees of conventionalization of linguistic units or structures. Conventionalization, however, is a process taking place first and foremost in social, rather than cognitive, systems, and it requires an additional logical step to assume that degrees of conventionalization more or less directly translate into degrees of entrenchment.

The corpus frequency of a phenomenon is hence due to the increased conventionalization of that phenomenon and in a second step it can be hypothesized to be related to the entrenchment of the phenomenon in the speakers' minds. Differences in corpus frequency can consequently be understood to reflect differences in conventionalization and degree of entrenchment. An experimental setting has been devised to systematically analyze how well conversion is established in the speech communities studied here and also in the speakers' minds. If the variation between the varieties of English which has been found in the corpus analysis is due to differences in the language system, these differences will resurface in an experiment.

In order to validate the findings from the corpus analysis, an experiment has been constructed. The goal of this experiment is to find out whether verb-to-noun conversion as a productive nominalization process is entrenched to different degrees in Asian and native varieties of English. Differences in conventionalization and entrenchment are expected to induce differences in how acceptable conversion is found to be and how it is processed. Firstly, higher degrees of conventionalization will prompt higher acceptability ratings. Secondly, a higher degree of entrenchment is hypothesized to result in higher processing speed, which can be measured through reaction times to a stimulus containing verb-to-noun conversion.

The experiment consists of two tasks and a background questionnaire. The first task is an acceptability rating task that assesses the degree of conventionalization of conversion in the speech communities corresponding to the varieties of English analyzed. The second task is a two-alternative forced-choice reaction time task (maze task) in which it is to be determined how well entrenched conversion is in the participants' minds.

8.1 Task 1: Rating task

In order to ascertain the degree of institutionalization of conversion, an acceptability judgment task (cf. section 4.3.1) has been chosen. The present task is a rating task in which speakers have to rate a sentence on a scale depending on how acceptable or unacceptable they find it. This task is also called *Likert scale task* (cf. Schütze and Sprouse 2013: 33), but the present task differs from the traditional Likert scale task as described in Schütze and Sprouse (ibid.: 33–34) in both the scale that is used and also in the instructions that the participants receive.

Schütze and Sprouse (ibid.: 33) identify the (usually) five- or seven-point Likert scale as the main disadvantage of this type of task. The fact that participants cannot choose an intermediate value could lead to uneven intervals between the points (because a value of a little under 2 and a value of a little over 2 both receive the score 2). Also, speakers might judge the difference between two (mathematically) identical intervals differently. Therefore, in this task, similar to Baroni et al. (2009: 47–48), who analyzed ratings of Italian deverbal nominal constructions, a near-continuous slider scale ranging from 0 to 1000 is used. (Perek and Hilpert (2014: 275), for example, prefer a scale with a color gradient over a discrete scale.) Figure 8.1 shows the slider scale. The red slider can be positioned freely on the scale. The number that corresponds to the position of the slider on the scale is not displayed. However, there are four labels next to the scale which help the participants in their judgment and which also serve the purpose of encouraging the participants to use the entire scale spectrum. (For the wording of the labels see below.)

This adapted scale has various advantages. First, participants are not biased by the numbers they see on the scale. Second, participants can easily rate sentences in a very fine-grained way. Third, the near-continuous scale (compared to a five- or seven-point Likert scale) is preferable for linear regression modeling, as this statistical method can only take continuous values as input (cf. Baroni et al. 2009: 47–48; Grace-Martin n.d.).

This rating task further differs from traditional rating tasks in that it refrains from explicitly asking participants for a metalinguistic judgment. Even though Schütze and Sprouse

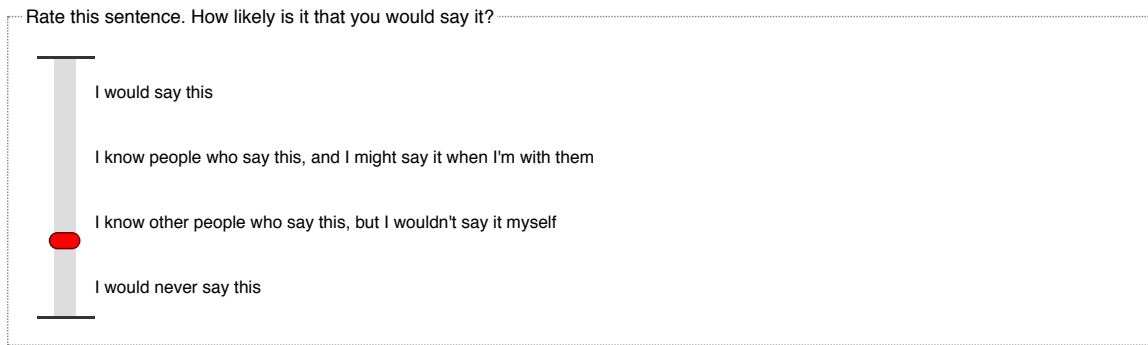


Figure 8.1: Slider scale used in the rating task

(2013: 28–29) make a point in dispelling doubts about acceptability judgment data, critical voices remain (cf. e.g. Bresnan 2007: 91, cf. section 4.3.1). In order to avoid the distortion of data by explicitly asking how ‘acceptable’ or ‘grammatical’ a sentence is deemed to be, the instructions for this task have been phrased in a way that asks the participants to identify with the target sentences. Along the scale there are four statements which indirectly describe the degree of acceptability of a sentence. These are:

- I would say this
- I know people who say this, and I might say it when I’m with them
- I know other people who say this, but I wouldn’t say it myself
- I would never say this

The more likely participants find it that they would produce the sentence in question themselves, the more acceptable the sentence is to them and the higher their rating will be.

8.2 Task 2: Maze task

The maze task (cf. Forster 2010; Forster et al. 2009) is a task that combines self-paced reading with forced choice while measuring reaction times. In self-paced reading tasks, participants themselves determine how fast stimuli are presented to them by clicking or pressing keys when they have read and understood what is displayed on the screen.¹ In forced-choice

¹For methodological challenges of standard self-paced reading tasks cf. Forster (2010: 347–350) and Forster et al. (2009: 163).

tasks, participants choose the most or least acceptable option out of two or more. In the present task, both methods are combined. Participants are to build sentences word for word, to metaphorically follow the sentences “through a maze” (Forster n.d.). At each step, two alternatives are presented and speakers have to choose the one that is more likely, i.e. more grammatical, at that particular position in the sentence by pressing the corresponding key on the keyboard. Participants’ reaction times are recorded for every decision. The more entrenched a construction is, the easier the retrieval process is going to be, and hence the less time participants will need to react to a stimulus containing this construction. To see how well verb-to-noun conversion is entrenched, participants are asked to construct sentences containing verb-to-noun converted forms.

Thus, the main feature of the maze task is that it “forces the reader into an incremental mode of processing in which each word must be fully integrated with the preceding context before the next word can be considered” (Forster et al. 2009: 163). The main advantage of incremental processing is that it enables the experimenter to gain information on the processing of each individual word (cf. Forster 2010: 350–351). Furthermore, the maze task offers other advantages that are of particular relevance in a web-based experimentation setting. First, since completing the task compels participants to understand the sentence, it is not necessary to check whether they have understood the sentence with post-task comprehension questions (ibid.; cf. Forster et al. 2009: 164). Second, as every step in the task requires that the words shown be “fully integrated with the preceding context”, there is no possibility for the participant to “adopt a ‘wait-and-see’ strategy” (ibid.: 163–164), i.e. to defer the decision process to a following word.² Third, “the response criterion is [...] clearly defined” (Forster 2010: 351): By design, there is always only one correct answer so that assessing whether a participant has chosen the right answer—and consequently also understood the sentence—is very easy for the experimenter. In short, the maze task demands a high level of commitment on the part of the participant while at the same time allowing for easy assessment of response accuracy for the experimenter. The maze task thus lends itself to a web setting, where participants cannot be monitored as they go through the experiment. Participants who show highly deviant reaction times or much higher than average error rates can easily be identified and their answers rejected.

Forster (ibid.: 352) proposes two versions of the maze task, the G-maze and the L-maze, where G stands for grammaticality and L for lexicality. In a G-maze, participants are presented with two actual words, one out of which is ungrammatical. In an L-maze, a word is

²In reading tasks, readers will often process words after they have read them and have moved on to the next word or words. This phenomenon is known as the *spillover effect*. By using the maze task, spillover effects can largely be avoided (cf. Forster et al. 2009: 164).

presented alongside a non-word, so that participants have to choose which of the alternatives is the ‘real’ word (also known under the name of lexical decision task). The L-maze offers an advantage over the G-maze in that it can mitigate the difference in results that is due to the “nature of the incorrect alternatives” (e.g. semantic aspects, Forster et al. 2009: 166). Additionally, the L-maze is recommendable if one wishes “to measure performance on semi-sentences of doubtful acceptability” (Forster 2010: 352). Verb-to-noun conversion is highly likely to constitute an instance of language use of “doubtful acceptability” to the participants speaking a native variety of English. Nevertheless, in this experiment the G-maze paradigm is chosen. In a G-maze task where the ‘correct’ alternative is not correct but only more likely, reaction times “primarily reflect response uncertainty” (Forster et al. 2009: 167). The following equations illustrate the reaction times (RT) for the standard G-maze and the modified G-maze paradigm.

$$\text{standard G-maze: RT} = T_{\text{processing}} + T_{\text{decision}} \quad (8.1)$$

$$\text{modified G-maze: RT} = T_{\text{processing}} + T_{\text{response uncertainty}} + T_{\text{decision}} \quad (8.2)$$

In a study on World Englishes, the modified G-maze paradigm is highly useful. In the present study, the target sentences each contain one instance of verb-to-noun conversion, so that participants are forced to choose a converted noun to complete the sentence. The converted form is expected to be deemed ungrammatical by some of the participants, depending on the variety they speak. The converted form is presented with an alternative that is even more ungrammatical (e.g. determiner following determiner: **a the*). In those cases, where the participants are more familiar with verb-to-noun conversion (e.g. HKE), participants will not notice that the two options presented are both ungrammatical in Standard English grammar. For this group, the task will be a standard G-maze task and choosing between a converted form and an alternative will be like the other choices and will hence be as readily made as the other choices. For speakers of those varieties where verb-to-noun conversion is infrequent (e.g. speakers of the native varieties), the process of choosing the ‘correct’ option is expected to take longer, since this particular step is the modified version of the G-maze task where reaction times are composed of the time it takes to process the two words and to make a decision, and also the time that reflects the response uncertainty. In this case, participants have to infer that they are to choose the least unacceptable option, which is the converted form. As this inference comes at a higher processing cost and since this is likely to cause uncertainty, the interval between the appearance of the word pair on the screen and the action of indicating the ‘correct’ option is expected to be of a longer duration than the rest of the choices.

Consequently, a difference in reaction times between speakers of new and native varieties is hypothesized to be due to the absence of decision uncertainty for the speakers of those varieties in which V>N conversion is comparatively frequent. While this task is not intended to provide information on the absolute time it takes participants to process words, it aims to establish whether there are differences in reaction latencies. These can then be traced back to response uncertainties resulting from different degrees of entrenchment of V>N conversion.

8.3 Task 3: Background questionnaire

The third task of the experiment is a short questionnaire on participants' language use and language learning background. The main independent variable in the two previous tasks is the variety of English which participants speak. The variety of English is closely related to the native language and the home country of the participants. Also, the degree to which English is used in daily life is considered to be of relevance. The background questionnaire thus seeks to obtain detailed information about the linguistic ecology in which participants find themselves. It is further used to gather other sociolinguistic variables that might influence acceptability judgment and reaction times (e.g. level of education, gender or age). The exact wording of the background questionnaire can be found in appendix E.3.

8.4 Hypotheses

In line with what has been outlined above and on the basis of the results of the corpus study, the following hypotheses can be formulated for the experiment:

1. Speakers of New Englishes and native varieties will find conversion acceptable to different degrees, with speakers of new varieties rating conversion as more acceptable. This is supposedly due to two mechanisms. The first is transfer from the substrate languages, the second is the development of new local norms in these varieties. Both mechanisms are potentially moderated by the sociolinguistic status of English in the respective areas.
2. Speakers of HKE will find conversion more acceptable, i.e. give it a higher rating, than speakers of SgE and IndE, reflecting the lower degree of institutionalization of English in Hong Kong. Furthermore, speakers of HKE will readily opt for V>N conversion

where there is no other option available (maze task). This behavior will prompt comparatively lower reaction times to verb-to-noun conversion than for speakers of any other variety. Lower reaction times are associated with faster processing, which is the consequence of higher levels of entrenchment of this construction, which in turn is caused by the increased experience that a speaker has with this construction. This experience can be hypothesized to result from the influence of the Chinese substratum.

3. In the more advanced Chinese-substratum variety, SgE, the higher degree of institutionalization is expected to moderate transfer from the substratum to a considerable degree, leading SgE speakers to being exposed to V>N conversion less often. This will trigger higher reaction times and lower acceptability ratings for verb-to-noun conversion compared to HKE speakers.
4. Differences between SgE and IndE speakers are likely to be due to the influence of substrata. The socio-institutional status of English is assumed to be of less importance in comparing SgE and IndE, considering that both varieties have reached the same phase in the Dynamic Model, that is, present a similar degree of institutionalization of English.
5. Speakers of USE and BrE will (only) show slight differences in behavior. Overall, in line with the corpus analytic findings, these two groups are estimated to have the least experience with V>N conversion, considering that the phenomenon occurs very infrequently in these varieties. It is therefore expected to be less entrenched, which will induce speakers of the native varieties to display the highest reaction times for verb-to-noun conversion as well as to rate verb-to-noun conversion the lowest.

8.5 Materials and design

In this section, the materials for the rating and the maze task as well as the background questionnaire are presented.

8.5.1 Rating task

The rating task aims to obtain judgments on the acceptability of verb-to-noun conversion in various varieties of English. Participants are presented with one sentence at a time and are

asked to rate how likely it is that they would produce this sentence themselves. The set of stimuli is composed of ten target stimuli, ten control stimuli and 30 distractor stimuli.³

Target stimuli

The group of target stimuli, that is, sentences with verb-to-noun conversions, is composed of sentences from the dataset obtained during the corpus analysis (cf. section 6.2) and further sentences from GloWbE that were obtained by a manual search for the verbs in the dataset in distinctly nominal contexts (preceded by a determiner). All sentences but one stem from the varieties analyzed; one sentence is from the New Zealand section of GloWbE. Only sentences with easily understandable content have been chosen, so as to avoid the influence of unknown lexical items on the rating result (cf. Schütze 1996: 185–186). Where necessary, the sentences have been modified in such a way that they show no other non-standard feature than a novel verb-to-noun conversion. Additionally, some sentences have been shortened so as to minimize the processing burden. The following sentence pair is an example of such modifications.

- (8.3) original: he saw some state troopers and was sure they'd arrest him for possess
of alcohol
modified: He was sure they'd arrest him for possess of alcohol.

Control stimuli

The group of control stimuli consists of sentences in which the converted nouns from the target stimuli are used as verbs. These sentences, contrary to all other sentences in the set, contain no non-standard features. For every target sentence, a control sentence from COCA or the US section of GloWbE has been matched. For all but two sentences, the utterance length of the target sentence is the same or differs in only one word. In two sentences, the length of the control sentence exceeds the length of the target sentence by two or three words. The control stimuli are once again chosen in such a way that the words used are easily understandable. The sentences in 8.4 are an example of a target stimulus and the matched control stimulus.

- (8.4) target: This post is the continue of my last post on May 10. (GloWbE-HK)
control: There's really no reason or excuse to continue with this. (COCA-SPOK)

³All stimuli for the rating task are listed in appendix E.1.

Distractors

The distractor items serve two functions. Not only do they distract from the target items (i.e. act as fillers), but they also serve to evaluate the overall rating behavior of the participants. There are three types of distractors in this task. They all contain non-standard features but are hypothesized to be rated differently, depending on the native variety of the rater. Whether features occur in the varieties was determined by drawing on the ratings in eWAVE (cf. Kortmann and Lunkenheimer 2013a).

The three types of distractor items are described in the following and illustrated in more detail in table 8.1.

non-standard feature, occurs in none of the varieties The first type of sentences includes non-standard features that are not used in either of the varieties analyzed according to the eWAVE database. These features have a rating of ‘D’ or ‘X’. All participants are expected to rate these sentences as unacceptable.

non-standard feature, occurs in select varieties The second type of sentences shows non-standard features that only appear in HKE and SgE, but not in IndE, BrE, and USE. These features have a rating of ‘A’ or ‘B’ in HKE and SgE and a rating of ‘D’ or ‘X’ in IndE, BrE, and USE.⁴ Participants who speak HKE or SgE are predicted to rate these sentences more favorably than participants with an IndE or a Standard English language background.

non-standard feature, occurs in all varieties The third type of sentences contains features that are non-standard in nature but still pervasive in all varieties, that is, have obtained a rating of ‘A’ or ‘B’ in all varieties investigated. Participants who rate these sentences as comparatively unacceptable can be assumed to possess a high language awareness.

Due to the subjectivity of the ratings in eWAVE (cf. Kortmann and Lunkenheimer 2013b), unclear ratings (‘B’ or ‘C’) or ratings that did not correspond to the literature (see above) were cross-checked in GloWbE. An example is the use of *there is* with a noun in the plural (feature #172). This feature has the rating ‘D’ in SgE. However, a search for this construction in GloWbE reveals that its frequency in SgE clearly ranges above the average frequency of this construction in GloWbE.

⁴A notable exception is feature #92, which is ranked ‘?’ in HKE and ‘D’ in SgE, but is listed as a highly pervasive feature of HKE in Setter et al. (2010: 55–56). A search in GloWbE reveals that the combination of *they* + verb in present tense, third person singular ([v?z*]) has a higher than average frequency in both HKE and SgE. This feature is therefore included despite its ranking in eWAVE.

Table 8.1: Distractor items in the rating task

type of distractor	example sentence	GloWbE section	eWAVE number
rated ‘D’ in all varieties	<i>If unu disagree with the message of the video, unu can do nothing about it.</i>	Jamaica	23
rated ‘A’ in HKE, SgE	<i>We \emptyset always looking for the best things for our business.</i>	HK	174
rated ‘A’ in all varieties	<i>I remember my mother and myself walking around the streets of Paris.</i>	HK	8

The distractor stimuli have been drawn from various sections of GloWbE. Where necessary, sentences have been modified so that lexis that reveals the origin of the sentence (e.g. Asian-sounding proper nouns) and which could therefore influence the rating has been replaced by neutral lexical items.

The distractor items are distributed among the three types as follows. There are 15 sentences with features that occur in none of the varieties, 10 sentences with features that occur in all varieties, and 5 sentences with features that occur only in HKE and SgE. This sums up to a total of 50 sentences in this task. Table 8.2 gives an overview of the stimuli. The entire list of stimuli for the rating task is shown in appendix E.1.

Table 8.2: Overview of the stimuli in the rating task

group	no. sentences	example sentence
I target: V>N conversion	10	<i>Vaccinations can help stop the distribute of viruses.</i>
II control: verbal use	10	<i>We can’t distribute them here in the country.</i>
III rated ‘D’ in all	15	<i>It’s one of them books where you don’t want to miss a thing!</i>
IV rated ‘A’ in HKE/SgE	5	<i>She lingers for as long as she can before she walk-\emptyset away.</i>
V rated ‘A’ in all	10	<i>If I was younger, it wouldn’t bother me.</i>

The fact that participants who speak HKE or SgE are expected to rate the five sentences that contain HKE/SgE-only features better than speakers of the native varieties introduces a small bias in the HKE and SgE group of participants. The speakers of HKE or SgE might

be slightly biased towards a positive rating (25 supposedly acceptable to 15 supposedly unacceptable sentences). For the speakers of BrE and USE, there is no bias (20 supposedly acceptable to 20 supposedly unacceptable sentences), following Schütze's (1996: 185–194) recommendations.

The sentences are presented to each participant in random order (cf. *ibid.*: 187). To rate a sentence, participants have to click on the slider bar for the red slider to appear. The slider can then be moved to the desired position by dragging-and-dropping. Figure E.1 in appendix E.4 is an example of what the rating question looks like for the participants.

8.5.2 Maze task

The aim of the maze task is to gain insight into how verb-to-noun conversion is processed. In this task, participants are asked to construct sentences word for word. The first word of each sentence is presented alone. All subsequent words in the sentence are each presented with an alternative word. The participants' task is to press the left or right arrow key, corresponding to the side on which the word that they find more likely/more grammatical is shown. In order to make the choice easy, structurally impossible or highly unlikely alternatives are given together with the correct word. The first word of each sentence is displayed on the screen after participants have started the trial; all subsequent word pairs appear when they have made their choice by pressing one of the arrow keys.

The set of stimuli consists of 30 target sentences and 30 filler sentences.⁵ Most target sentences that the participants are to assemble have been taken from the dataset gathered for the corpus analysis (cf. section 6.2). Further sentences have been obtained from GloWbE by searching for further potentially converted verbs in distinctly nominal contexts. The potentially converted forms have been taken from the original list of verbs that have no corresponding nominal form according to the OED (cf. section 6.2). The nominal contexts that were searched were the verb preceded by a determiner, the verb preceded by a determiner and an adjective, and the verb preceded by a determiner, an adverb and an adjective.⁶ Singular and plural forms of the converted nouns were input in the search.

Where necessary, the sentences obtained from GloWbE have been modified so as to show no other non-standard features besides the converted form. Sentences have further been edited to contain no ambiguous adjective-noun combinations where the adjective could be understood as the (converted) head of a noun phrase and the converted noun as the main verb of the clause (e.g. *the most interesting discovers* > *all these interesting discovers*). In some cases,

⁵The complete list of stimuli is available in appendix E.2.

⁶The search queries in GloWbE were [at*] V, [at*] [j*] V, [at*] [r*] [j*] V, respectively.

sentences have been shortened by deleting subordinate clauses or omitting coordinated parts of phrases. An example of such manipulations of the original sentence can be seen in 8.5. An example of a stimulus, consisting of the target sentence (on the left) and the ungrammatical or highly unlikely alternatives (on the right), is given in table 8.3.

- (8.5) original: match outcome is very possible beyond our expect
 modified: The outcome of the match is very possibly beyond our expect.

For the control sentences, the target sentences have been modified in such a way that they contain derived nouns instead of converted nouns. The target and control sentences are distributed across two lists, so that each list comprises 15 target sentences with converted nouns and 15 control sentences with derived nouns. The participants are randomly assigned to either of the two lists. The reaction times to the derived nouns serve as a basis of comparison for the reaction times to the converted nouns. In neither of the sentences does the target word, that is, the converted or the derived form, occur more than once, so as to avoid priming effects.

Table 8.3: Example stimulus for the maze task

The	
doctor	therefore
told	fifth
him	boil
that	drawer
he	went
had	they
only	market
one	must
choose.	every.

Participants start each round by pressing one of the two arrow keys once they see a ‘Start’ sign presented in the middle of the screen. The first word of the sentence is then displayed in the middle of the screen. After pressing either arrow key for the first word, the next word and the ungrammatical/less grammatical alternative are presented in two fields, one on the left and the other on the right. The side on which the words appear is randomized for each participant. Participants then press either the right or left arrow key to indicate their choice. Figure E.2 illustrates what the maze task looks like for the participants.

For all choices, reaction time is measured in milliseconds. Reaction time is operationalized as the time interval between two key presses, i.e. two choices, or the first pressing of the arrow key to start the next round and the first choice.

The original procedure proposed by Forster et al. (2009) is slightly modified in this study. Contrary to what Forster et al. (ibid.: 164) propose, a sentence is not aborted when participants make a mistake, and participants do not receive feedback on whether their answers are correct or not. Aborting sentences could encourage participants to carelessly choose any of the alternatives merely to end the experiment quickly. Furthermore, since the task is highly complex, it is expected that participants make mistakes frequently. In order to avoid demotivating the participants, no feedback is provided.

8.5.3 Background questionnaire

The background questionnaire is a digital form into which details can be entered by typing or by clicking on radio buttons. The background questionnaire as seen by the participant is shown in figure E.3 in appendix E.4.

8.6 Procedure

The experiment was programmed using the software *QualityCrowd2*, originally designed for video quality assessment tasks (cf. Keimel et al. 2012).⁷ There are various benefits to using the *QualityCrowd2* software for this experiment. First, it is compatible with various crowdsourcing platforms and can also be used for ‘simple’ web-based experiments, which allows launching the experiment on various platforms without changing the script. If one were to repeat the experiment, for example with participants of other varieties or on a crowdsourcing platform, the code could simply be reused. Second, scripting the experiment is comparatively easy with this software. Also, *QualityCrowd2* is free, open-source software, which simplifies adapting tasks to specific requirements of the research project or adding new task formats such as the maze task to the list of available types of tasks. Furthermore, one of the tasks already available in the *QualityCrowd2* software is a rating task with a continuous slider.

The order of the tasks in the experiment was the same for all participants. After the introductory text, sociolinguistic variables were gathered by means of the background ques-

⁷The version that was used for programming and running this experiment is newer than the one mentioned in Keimel et al. (2012). It can be downloaded from <https://github.com/clorch/QualityCrowd2> (Horch 2015).

tionnaire.⁸ The reasons for asking for personal information at the beginning of a web-based experiment have been laid out in section 4.3.3. The next task is the rating task, followed by the maze task. The maze task was put last since the maze task might lead participants to think that the study is about conversion, which would influence the rating were it to succeed the maze task.

The link to the experiment was posted on Facebook⁹ and distributed via e-mail. The researcher asked friends and colleagues to share the link to the experiment with their friends and students¹⁰ and motivate them to distribute the link even further. The link was also posted to various ‘Facebook groups’¹¹ and mailing lists (e.g. LinguistList¹²).

In order to partly emulate a crowdsourcing environment, financial incentives were provided. Paying every single participant individually would have been unfeasible due to time constraints, so that a raffle scheme was adopted in which thirty payments of 15 euros (totaling 450 euros) were given to randomly selected participants. The payments were sent through PayPal¹³.

8.7 Pre-test

A pre-test with 18 native speakers of German was conducted to test the procedure and measure the time required for the experiment. All participants have received instruction of the English language in school and can be assumed to have a high to very high knowledge of English. Subjects did not receive financial compensation for their participation.

The link to the pre-test was sent out by the researcher to friends and acquaintances. One participant’s data had to be deleted because the total experiment time was two standard deviations above the mean. Out of the remaining 17 participants who finished the experiment, 10 were female and 7 were male, with an average age of 25.3 years (range 17 to 54 years, median 22.0 years). The mean experiment duration was 31.5 minutes.

⁸In the pre-test, the background questionnaire was located at the end of the experiment. Considering that all participants in the pre-test know the researcher personally, it was considered unnecessary to implement the high-hurdle technique (cf. section 4.3.3).

⁹<https://www.facebook.com>

¹⁰Asking linguists to distribute the link among their students is not part of the friend-of-a-friend approach as outlined in section 4.3.3, however, this step had to be taken in order to maximize the spread of the call.

¹¹Facebook users can ‘join’ groups to connect with other users who share similar interests (e.g. music, hobbies) or qualities (e.g. have all received a scholarship from the same organization) or life trajectories (e.g. have all emigrated from their home country to another country).

¹²<http://linguistlist.org>

¹³<https://www.paypal.com>

Linear regression was used to analyze the results. In the rating task the sentences containing features rated ‘D’ were rated highly significantly worse ($p < .001$), and sentences containing features typical of HKE and SgE were rated significantly worse ($p < .01$) than the other types of sentences. The fact that the German native speakers did not assign significantly lower ratings to the target sentences than the control sentences can be hypothesized to be due to the participants’ insecurity in English usage. Participants’ metadata (age, gender, education) were not considered in this analysis. The result of the linear regression model for the data from the maze task revealed that the German participants showed significantly higher reaction times for the converted forms. This is in line with the fact that German high schools in teaching English orient towards the British or American English standard, in which V>N conversion is highly constrained, leading to participants’ unfamiliarity with the innovative forms.

At the end of the experiment, participants were asked to report their experiences. Some participants also directly reported to the researcher. The test was then slightly modified on the basis of the participants’ feedback. The most important aspect is that participants almost unanimously reported difficulties in understanding the maze task. Therefore, two trial sentences were integrated so as to train participants before exposing them to the actual test items. Another change was the inclusion of a warning that the experiment cannot be done on devices without a keyboard (e.g. smartphones, tablet computers).

8.8 Participants

The experiment was run from July 15th, 2015, to November 15th, 2015. Potential participants were contacted over Facebook and via e-mail. This was to predominantly recruit (regular) users of the internet, who can be assumed to overlap at least in part with the authors of the texts represented in GloWbE. As Schnell et al. (2011: 377) point out, the resulting sample is a “convenience sample”, that is, it is not necessarily representative of the population. This means that all generalizations and conclusions drawn on the basis of the data obtained from this sample have to be interpreted very carefully.

During the time period the experiment was online, the link to the experiment received 1226 clicks. A total of 208 participants (17%) completed the experiment. Of these, 25 indicated either a country of residence or growing up or a native language that did not fulfill the prerequisites for the experiment, so that their answers were discarded. The average duration as measured by the median of the time it took participants to complete the experiment was 24.2 minutes. The mean (293.3 minutes) was highly skewed due to various outliers with ex-

tremely high completion times. Therefore, a threshold beyond which a participant's answers were considered unreliable was manually set to 70 minutes. In analogy, a total completion time of below 10 minutes was considered unreliable, so that answers by participants with completion times below 10 minutes were rejected as well. This resulted in the removal of data by 11 participants. After adjusting the completion times, the median is 24.02 and the mean is 25.75 minutes. Figure E.4 in appendix E.5 visualizes the amount of time participants needed to complete the experiment.

After deleting answers by participants with unreliably short or long completion times, the total number of people who participated in the experiment amounts to 172 (14% of 1226). The distribution per variety and per list is represented in table 8.4.¹⁴ Of these participants, 100 are female, 71 are male, and 1 classified themselves as belonging to another gender. The mean age is 27.9 years (minimum 16 years, maximum 71 years), with the mean age being comparatively similar across varieties. That fact that the mean age lies between twenty and thirty years is of little surprise considering that the friend-of-a-friend approach was adopted to recruit participants. The age of the participants is expected to be similar to the researcher's age as well as the researchers' friends' age. The educational background of the participants is mostly academic, as can be expected on the basis of the recruiting procedure (friend of a friend of an academic). Across varieties, the educational background varies slightly, which can be assumed to be due to the correlation of age and education. As figure E.6 in appendix E.6 shows, BrE participants have a comparatively lower level of education, which correlates with the lower average age of this group (cf. table E.3 in appendix E.6). More detailed descriptions of participants' gender, age, and educational background can be found in appendix E.6.

Table 8.4: Participants per variety and list

variety	list 1	list 2	total
USE	24	35	59
BrE	8	24	32
HKE	9	7	16
IndE	17	25	42
SgE	14	9	23
total	72	100	172

¹⁴As can be observed on the basis of table 8.4, participants are not evenly distributed across varieties. Unequal group sizes do generally not impact the feasibility of fitting a regression model (cf. Slinker and Glantz 1988: 354), as long as the number of observations per group is not so low as to impede the converging of the model.

As for the participants' linguistic background, USE and BrE offer a highly similar picture. Almost all participants are monolingual¹⁵ English speakers. For the Asian varieties, the results largely correspond to what has been outlined in section 1.1.3. Only very few of the SgE participants reported a language other than English as their native language, with the majority of participants indicating that they are English-Chinese bilinguals. This reflects the findings by e.g. Tan (2014) and Schneider (2014a: 250–251), who observe that English, even though not an official mother tongue in Singapore, is in actual fact the native language of many Singaporeans. In contrast, extensive bilingualism—as witnessed in Singapore—is the exception rather than the norm for the participants from Hong Kong. Most of them claim to have a dialect of Chinese as their native language and only a few are English-Chinese bilinguals. None of the HKE participants reported to exclusively have English as their native language. For IndE, the picture is similar with only very few monolingual English participants. The distribution of native languages per variety is displayed in figure 8.2. Among the native languages of those participants who gave neither English nor a Chinese dialect nor a combination of English and Chinese as their native language(s) are Hindi (13 participants), Tamil (6), Marathi (4), Malayalam (4), Kannada (3), Telugu (3), Bahasa Indonesia (1), Gujarati (1), and Tulu (1).

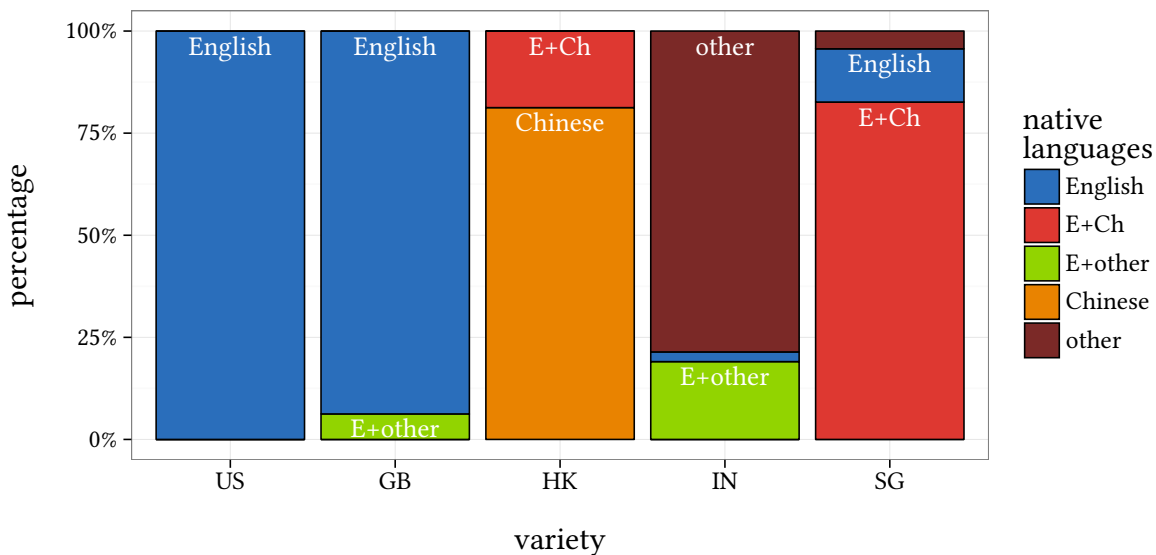


Figure 8.2: Native languages of participants per variety

Most participants use English on a daily basis, as figure 8.3 shows. However, 44% of all participants from Hong Kong and 19% of participants from India do not use English every

¹⁵Bilingualism in this context is used to refer to the acquisition of two languages from birth. Languages which participants reported to have learnt at a later stage, e.g. in school, were not considered.

day. This fact can be expected to be reflected in the English language proficiency of these participants.

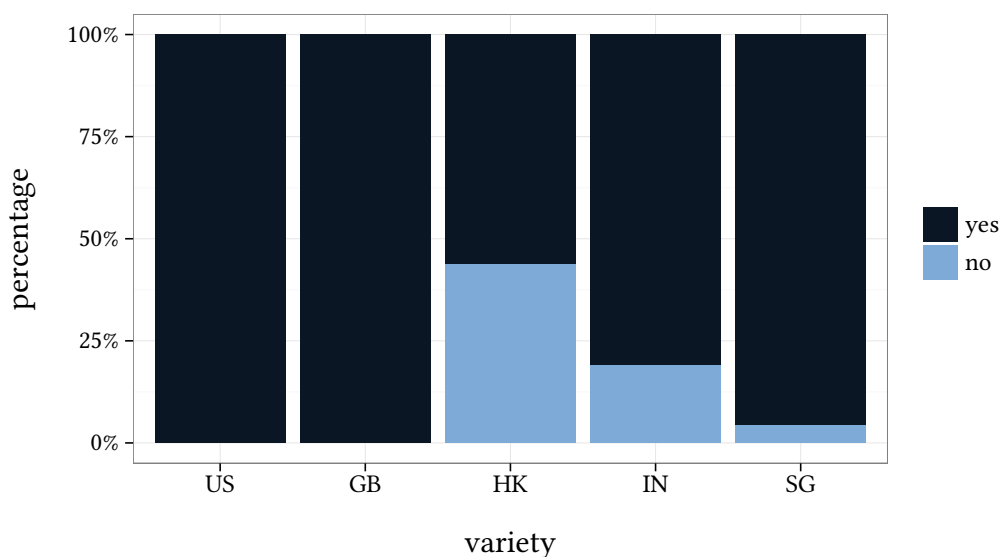


Figure 8.3: Participants' daily use of English per variety

8.9 Results

This section presents the results for the rating task and for the maze task separately. The results of both tasks were analyzed fitting linear regression models with random effects as described in section 4.2. An overview of the predictor variables common to both analyses and an explanation of them is given in the following. The levels set as the reference levels are indicated in parentheses.

age Age in years is a numeric variable, which was centered previously to analysis (cf. section 4.2.5). Younger speakers are expected to show more innovative language use (cf. Krug 1998: 180–182), potentially inducing a higher rating of sentences containing non-standard features. Furthermore, younger speakers are anticipated to react faster to stimuli, leading to shorter reaction times (cf. section 4.3.2).

gender (Female) The levels of this categorical variable are female, male, and other. Female speakers generally exhibit more innovative linguistic behavior (cf. Cheshire 2004: 429, Labov 2001: 501), which could potentially result in a higher rating of sentences with non-standard features. (For a comprehensive overview of the effects of gender on language variation cf. Cheshire 2004.)

education (Master's degree or higher) This categorical variable has been split into four levels of education: high school with no degree, high school diploma, bachelor's degree, and master's degree or higher. It can be assumed that the higher the education which the participant has received, the more norm-conforming their linguistic behavior will be. Conversely, less educated participants are predicted to accept the innovative form more readily (cf. e.g. Krug 1998: 179–180) and react to it more quickly. This variable could correlate with the age of the participant, since younger participants are presumably still in the process of pursuing a (university) career.

background in linguistics (No) Since the experiment was distributed via the friend-of-a-friend approach, it was considered necessary to record whether a participant's field of study or their professional occupation is related to linguistics or not. A background in linguistics is assumed to trigger a higher language awareness, potentially skewing the results of the rating task. Furthermore, a background in linguistics can result in participants uncovering the aims of the study more quickly, which would provide a distorted picture.

variety (US) The variety which participants speak was deduced from their country of residence and from the country in which participants indicated that they had grown up. Responses by participants with non-congruent answers (e.g. participant comes from Singapore, but grew up in Indonesia) were discarded. Variety is a categorical variable with five levels representing the five varieties investigated: HK, SG, IN, GB, US. For the same reasons which have been given in section 6.3.2, USE is set as the reference level.

daily use of English (Yes) This binary variable (yes, no) records whether participants use English every day. Irregular use of the English language can be hypothesized to result in a lower language proficiency.

L1 (English) Participants were asked to indicate their native language as well as list the three languages in which they are most fluent. Additionally, they were asked to give the number of years that they have been learning these languages. On the basis of this information, the participants' native language(s) were deduced and categorized into one of five groups.¹⁶ The five categories include English, a Chinese dialect, another native language, English and a Chinese dialect, English and another language. For the last two categories, only bilingualism from birth on was considered. Since this variable

¹⁶Particularly in the Singaporean context, some participants gave their official 'mother tongue' as their native language but reported that they had been learning both this language and English from birth.

correlates very strongly with the variety variable (collinearity)—almost all speakers of the native varieties are English monolinguals—it was not included in the analysis.

age of English onset The age of English onset was calculated as the difference between the age or the number of years learning the native language and the number of years learning English. However, not all participants provided enough information to include this variable as a predictor in the regression models. In a potential follow-up study, this variable would be of particular interest, considering that Chan et al. (2008: 34–35) and Yang et al. (2011: 670–680) found that early and late English-Chinese bilinguals process nouns and verbs differently.

random effect: Worker ID In both analyses, the individual participant (identified by means of a unique identification number called Worker ID in accordance with the terminology of Amazon’s Mechanical Turk) is included in the regression model as a random effect. This way, all effects that are due to idiosyncratic behavior are excluded from influencing the estimates of the other predictors.

8.9.1 Rating task

For the rating task, there are two further task-specific predictors, which are the type of sentence and the ID of the sentence.

type of sentence (Ctrl) As has been pointed out above (cf. section 8.1), five different types of sentences were presented in the rating task. These were sentences containing non-standard features with an eWAVE rating of ‘A’ in all varieties (A), sentences containing non-standard features with an eWAVE rating of ‘D’ in all varieties (D), sentences with non-standard features with a rating of ‘A’ only in HKE and SgE (AAsian), sentences containing V>N conversions (Target), and control sentences with no non-standard features, in which the target form is used as a verb (Ctrl). Depending on their native variety and on whether participants use English every day, they are expected to rate these groups of stimuli very differently. In order to account for this cross-varietal effect, interactions of variety and type of sentence and daily use of English and type of sentence are included in the model.

random effect: sentence ID In both tasks, trends are hypothesized to emerge independently of individual sentences which participants rate or build. The frequency of individual words or constructions which occur in the sentences are of crucial relevance to the

outcome of the rating. Therefore, in order to avoid that these frequencies influence the result, the individual sentences are added to the regression models as a random effect.

Across varieties participants made use of the entire rating scale. However, as can be seen in figure 8.4, participants oriented towards the four sentences describing the degrees of identification and often positioned the slider next to them, thus producing four peaks.¹⁷ Table 8.5 gives an impression of what the coded data look like.

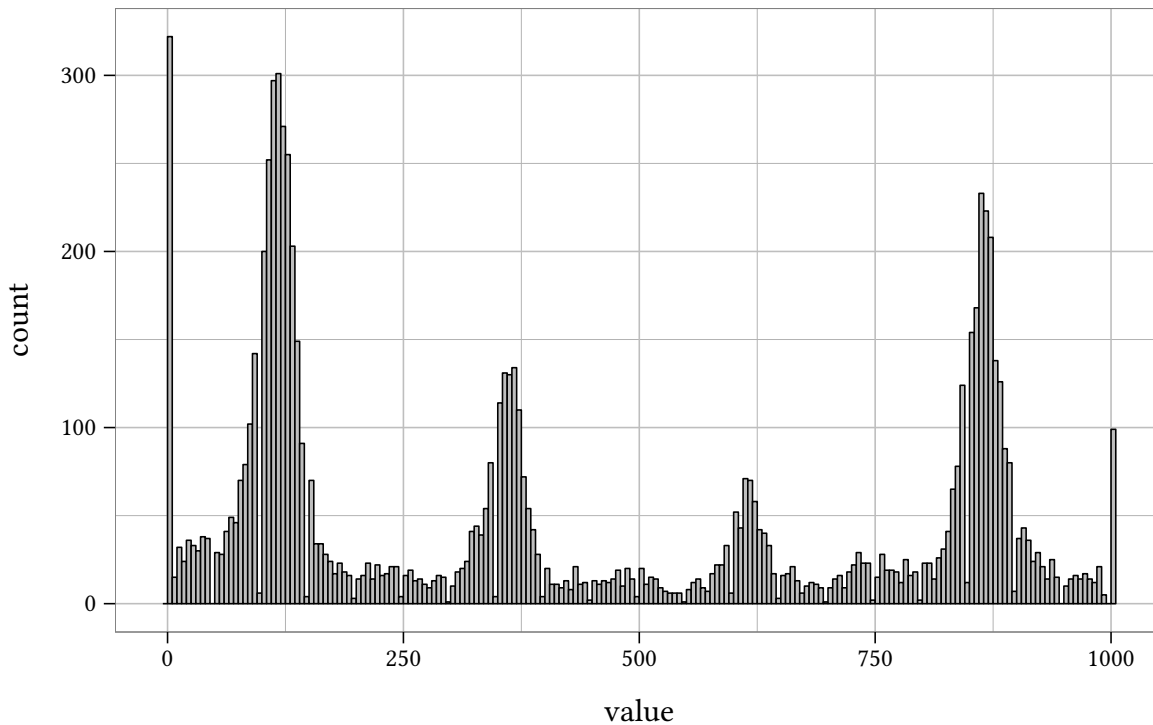


Figure 8.4: Histogram of the use of the rating scale (binwidth = 5)

A linear regression model including all of the above-mentioned predictors was fitted using the `lmer()` function. The formula is represented in 8.6.

$$\begin{aligned}
 \text{rating value} \sim & \text{variety} + \text{type of sentence} + \text{daily use of English} \\
 & + \text{linguistics background} + \text{age} + \text{gender} + \text{education} \\
 & + \text{variety} : \text{type of sentence} \\
 & + \text{daily use of English} : \text{type of sentence} \\
 & + (1 \mid \text{Worker ID}) + (1 \mid \text{sentence ID})
 \end{aligned} \tag{8.6}$$

¹⁷Figure 8.4 is a histogram. In order to create a histogram, the values of the independent variable (on the x-axis, in this case from 0 to 1000) are grouped into ranges of (usually) equal size, so-called bins. The binwidth (in this case 5) indicates the range. The values on the y-axis then show how many observations in a dataset fall into each bin (cf. Baayen 2008: 21–24; Field et al. 2012: 19).

Table 8.5: Subset of the coded data

WorkerID	sentencelD	value	typeSentence	variety	dailyUse	linguist	age	gender	education
1	55a64728c7c3ea	648	Target	US	yes	no	24.08	male	high school, no degree
2	55a64728c7c3ea	142	Target	US	yes	no	24.08	male	high school, no degree
3	55a64728c7c3ea	135	Target	US	yes	no	24.08	male	high school, no degree
4	55a64728c7c3ea	121	Target	US	yes	no	24.08	male	high school, no degree
5	55a64728c7c3ea	95	Target	US	yes	no	24.08	male	high school, no degree
6	55a64728c7c3ea	84	Target	US	yes	no	24.08	male	high school, no degree
7	55a654f5ac922a	117	D	US	yes	no	5.08	female	bachelor's degree
8	55a654f5ac922a	861	D	US	yes	no	5.08	female	bachelor's degree
9	55a654f5ac922a	850	D	US	yes	no	5.08	female	bachelor's degree
10	55a654f5ac922a	861	D	US	yes	no	5.08	female	bachelor's degree
11	55a654f5ac922a	128	D	US	yes	no	5.08	female	bachelor's degree
12	55a654f5ac922a	111	D	US	yes	no	5.08	female	bachelor's degree
13	55a654f5ac922a	122	D	US	yes	no	5.08	female	bachelor's degree
14	55a654f5ac922a	117	D	US	yes	no	5.08	female	bachelor's degree
15	55a654f5ac922a	122	D	US	yes	no	5.08	female	bachelor's degree
16	55a654f5ac922a	350	D	US	yes	no	5.08	female	bachelor's degree
17	55a654f5ac922a	122	D	US	yes	no	5.08	female	bachelor's degree
18	55a654f5ac922a	111	D	US	yes	no	5.08	female	bachelor's degree
19	55a654f5ac922a	117	D	US	yes	no	5.08	female	bachelor's degree
20	55a654f5ac922a	361	D	US	yes	no	5.08	female	bachelor's degree
21	55a654f5ac922a	117	D	US	yes	no	5.08	female	bachelor's degree
22	55a654f5ac922a	844	A	US	yes	no	5.08	female	bachelor's degree
23	55a654f5ac922a	856	A	US	yes	no	5.08	female	bachelor's degree
24	55a654f5ac922a	100	A	US	yes	no	5.08	female	bachelor's degree
25	55a654f5ac922a	850	A	US	yes	no	5.08	female	bachelor's degree
26	55a654f5ac922a	867	A	US	yes	no	5.08	female	bachelor's degree
27	55a654f5ac922a	861	A	US	yes	no	5.08	female	bachelor's degree
28	55a654f5ac922a	856	A	US	yes	no	5.08	female	bachelor's degree
29	55a654f5ac922a	861	A	US	yes	no	5.08	female	bachelor's degree
30	55a654f5ac922a	861	A	US	yes	no	5.08	female	bachelor's degree

However, age, gender, education, and linguistics background proved to be non-significant predictors, so that they were removed, yielding the final model the formula of which is reproduced in 8.7.

$$\begin{aligned} \text{rating value} &\sim \text{variety} + \text{type of sentence} + \text{daily use of English} \\ &+ \text{variety} : \text{type of sentence} \\ &+ \text{daily use of English} : \text{type of sentence} \\ &+ (1 \mid \text{Worker ID}) + (1 \mid \text{sentence ID}) \end{aligned} \quad (8.7)$$

A comparison of information criteria (AIC, BIC) for the full model (cf. 8.6) and the final model showed that the full model was not significantly better than the final model (without age, gender, education and linguistics background as predictors).¹⁸ The coefficients for the final linear regression model are displayed in table 8.6.¹⁹ The scaled residuals and the random effects are reported in table E.5 in appendix E.7.

The boxplots in figure 8.5 offer a visualization of the results (cf. Baayen 2008: 30). The ratings are presented separately for each type of sentence, and the values for every variety are given in a separate box (same color of box means same variety). The size of the box depends on the values themselves as the box comprises the values between the first and the third quartile²⁰ (interquartile range). The lines below and above the boxes (so-called whiskers) indicate “maximally 1.5 times the interquartile range” (ibid.). Values that fall outside 1.5 times the interquartile range are displayed as dots. These values are often called outliers but need not necessarily be removed from the data in order for the model to make sense. In this case, the high number of dots for USE for the control and the target sentences merely indicates that the USE participants provided fairly homogeneous ratings for these types of sentences whereas speakers of other varieties did not (e.g. HKE). The horizontal line in each box represents the median, the diamond the mean.

¹⁸AIC: 118,226.1 (full model) vs. 118,267.8 (final model), BIC: 118,508.5 (full model) vs. 118,500.7 (final model).

¹⁹Trimming the dataset as is done below for the maze task data, following Baayen (2008: 257–258), results in a model that is not significantly different from the model presented here. The only predictor that changes is daily use of English. Of the interactions between daily use of English and the type of sentence only the one between the type ‘AAsian’ and no daily use remains significant. This shows that the effect of the target sentences could be due to a small number of extreme rating values. However, this model is not reported in detail here, as the mere existence of extreme values does not necessarily justify this comparatively radical trimming procedure (cf. Osborne and Overbay 2004). Moreover, the residuals of the model for the rating task are almost normally distributed, in contrast to the first model fitted to the maze task data.

²⁰Dividing a set of data values into quartiles means dividing it into four groups of equal size. The three points that separate the four groups from one another are called quartiles.

Table 8.6: Rating Task

	Estimate	Std. Error	df	t val	p	
intercept (Intercept)	784.49	33.44	64.07	23.46	0.000	***
varieties						
varietyGB	-58.64	24.72	372.67	-2.37	0.018	*
varietyHK	-136.43	34.93	372.67	-3.91	0.000	***
varietyIN	-97.65	23.60	372.67	-4.14	0.000	***
varietySG	-118.06	27.71	372.67	-4.26	0.000	***
type of sentence						
typeSentenceA	-136.35	44.56	50.86	-3.06	0.004	**
typeSentenceAAsian	-504.06	54.58	50.86	-9.24	0.000	***
typeSentenceD	-521.15	40.68	50.86	-12.81	0.000	***
typeSentenceTarget	-612.63	44.56	50.86	-13.75	0.000	***
daily use of English						
dailyUseno	3.99	33.39	372.67	0.12	0.905	
variety : type of sentence						
GB : typeSentenceA	57.41	22.59	8361.94	2.54	0.011	*
HK : typeSentenceA	22.96	31.93	8361.94	0.72	0.472	
IN : typeSentenceA	0.03	21.57	8361.94	0.00	0.999	
SG : typeSentenceA	27.60	25.33	8361.94	1.09	0.276	
GB : typeSentenceAAsian	-1.59	27.67	8361.94	-0.06	0.954	
HK : typeSentenceAAsian	279.23	39.11	8361.94	7.14	0.000	***
IN : typeSentenceAAsian	189.45	26.42	8361.94	7.17	0.000	***
SG : typeSentenceAAsian	180.71	31.03	8361.94	5.82	0.000	***
GB : typeSentenceD	98.99	20.63	8361.94	4.80	0.000	***
HK : typeSentenceD	182.50	29.15	8361.94	6.26	0.000	***
IN : typeSentenceD	122.34	19.70	8361.94	6.21	0.000	***
SG : typeSentenceD	110.07	23.13	8361.94	4.76	0.000	***
GB : typeSentenceTarget	105.34	22.59	8361.94	4.66	0.000	***
HK : typeSentenceTarget	300.55	31.93	8361.94	9.41	0.000	***
IN : typeSentenceTarget	206.31	21.57	8361.94	9.56	0.000	***
SG : typeSentenceTarget	194.64	25.33	8361.94	7.68	0.000	***
type of sentence : daily use						
typeSentenceA : dailyUseNo	73.50	30.52	8361.94	2.41	0.016	*
typeSentenceAAsian : dailyUseNo	107.74	37.38	8361.94	2.88	0.004	**
typeSentenceD : dailyUseNo	38.04	27.86	8361.94	1.37	0.172	
typeSentenceTarget : dailyUseNo	67.34	30.52	8361.94	2.21	0.027	*

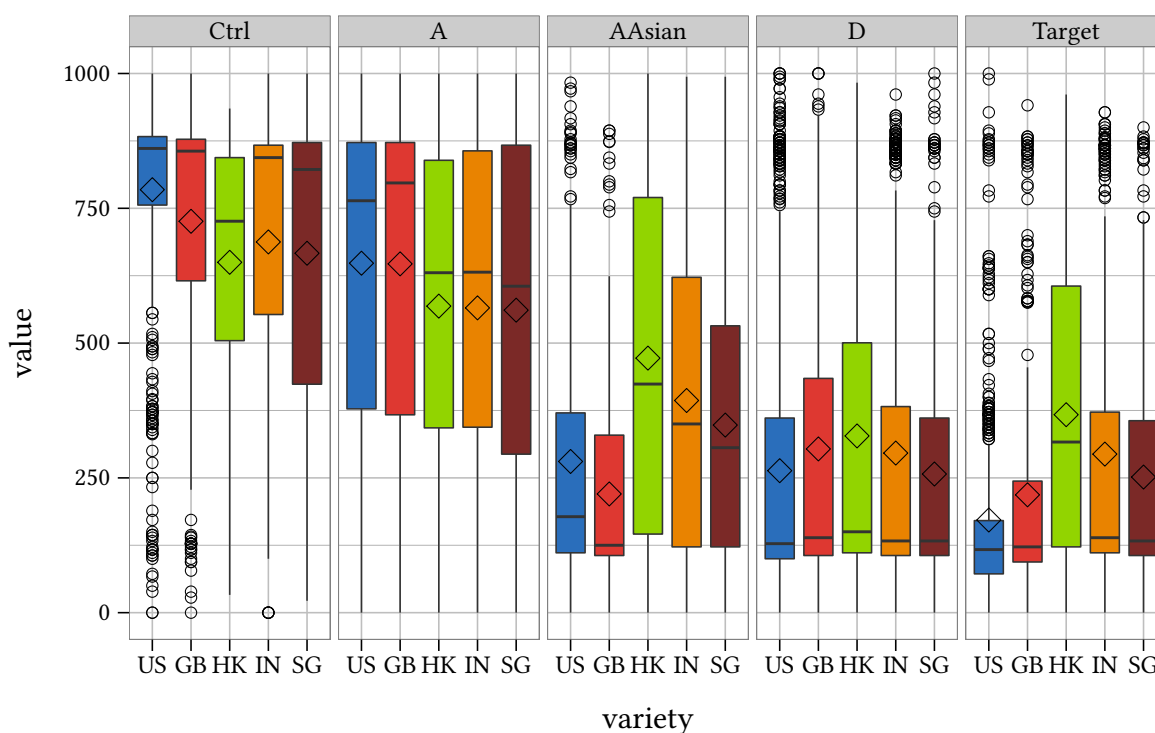


Figure 8.5: Boxplots of ratings per variety and per type of sentence. For each box, the horizontal line represents the median and the diamond represents the mean.

Results

Control sentences judged less acceptable in new varieties

In the output of the regression model as given in table 8.6, the intercept represents the mean rating for the control sentences by speakers of USE ($B = 784$). All other values in the table refer to this baseline value. Under the heading of ‘varieties’, the ratings for the control sentences (‘Ctrl’) in the other varieties are given.²¹ Overall, the rating of the control stimuli is significantly lower in all varieties compared to USE, but particularly low in the new varieties. This could be due to general insecurities of speakers in nativizing settings, where exonorma-

²¹As has been mentioned in section 4.2.2, the models presented use treatment contrasts. This means that all estimates for the levels of a categorical variable (such as variety or type of sentence) are given with reference to the pre-specified reference level (in the section ‘varieties’: variety = USE; cf. Levshina 2015: 185). When a categorical variable is involved in an interaction, “the interacting terms shown in the table [in this model: variety and type of sentence] are no longer main effects. They represent the estimates for the combinations of the specified level [in the section ‘varieties’: GB, HK, IN, SG] with the reference level of the interacting variable [in the section ‘varieties’: typeSentence = Ctrl]” (ibid.: 195). Thus, in the section ‘varieties’, the reference level of the interacting variable is the control sentences, and in the section ‘type of sentence’ the reference level of the interacting variable is USE.

tive standards are gradually being replaced by new local standards, but these endonormative standards have not been fully established yet.

Sentences with non-standard features rated worse than control sentences

The next block, under the heading of ‘type of sentence’, gives the ratings of all sentence types for USE, the reference variety. All other sentence types are rated significantly lower than the control sentences by speakers of USE. The effect is smallest for the sentences containing highly pervasive non-standard features (rated ‘A’ in eWAVE; $B = -136$). Sentences with non-standard features that do not occur in USE (types AAsian, D, and Target) receive a rating which is significantly lower, with the converted forms being rated the lowest ($B = -613$). This pattern varies drastically across varieties. In the following, each type of sentence is considered separately (reflecting the fifth section under the heading of ‘variety : type of sentence’).

Highly pervasive non-standard features not rated differently across varieties

For the sentences with highly pervasive non-standard features (with a rating of ‘A’ in eWAVE) the differences between the varieties that also apply in the case of the control sentences remain. That is, speakers of the new varieties rate these sentences comparatively lower than speakers of the native varieties. A small exception are speakers of BrE, who rate these sentences slightly better ($B = 57, p < .02$) than the control sentences compared to the USE speakers. However, this effect is rather marginal compared to the other effects present in the data.

Sentences with features highly pervasive in HKE and SgE receive higher ratings in all New Englishes

The sentences with features that are rated highly pervasive (‘A’) in HKE and SgE (short: ‘AAsian’) served the purpose of assessing in how far the raters participating in the experiment are typical representatives of their varieties.²² In general, the sentences containing non-standard features widespread in HKE and SgE receive worse ratings compared to the control sentences ($B = -504$ for the reference level), yet, this effect is not the same across all varieties investigated. As could be expected, the participants representing the native varieties rate these sentences significantly lower than participants from Hong Kong, India, and

²²Considering the subjective nature of the ratings in eWAVE, the results of this comparison of the rating data and the eWAVE data has to be interpreted with caution.

Singapore, and also do not differ significantly in their rating ($p < .96$, except for the small difference that is also present for the control sentences). Speakers of new varieties are more tolerant, and for HKE participants the size of this effect is largest ($B = 279$, $p < .001$). This is in line with what was hypothesized.

SgE participants rate these sentences significantly better than speakers of native varieties but do not rate these sentences considerably better than speakers of IndE, as a comparison of the sizes of this effect for these two varieties reveals ($B = 181$ for SgE vs. $B = 189$ for IndE). Considering that the features presented in these sentences are supposedly highly pervasive in SgE, they receive a rather low rating score (see below for discussion).

The fact that IndE participants should rate these sentences so highly seems unexpected, considering that these features are not as pervasive in IndE as in HKE or SgE. Nonetheless, as a search in GloWbE shows, all of these features also occur in IndE, although often to a smaller extent than in HKE and SgE. This could explain why IndE speakers still find the sentences containing these features fairly acceptable. What is more, this finding could be the consequence of a more general proficiency effect, which leads speakers of IndE to show insecurities in English language use, which then reflect in the higher acceptability of the sentences with features typical of Chinese-substratum varieties (see below).

Sentences with ‘foreign’ non-standard features receive worst ratings in USE and low ratings across all other varieties

The sentences containing features that are attested in neither of the varieties analyzed (‘D’) are generally rated lower than the control sentences (effect size for the reference level: $B = -521$). However, speakers of BrE, HKE, SgE, and IndE rate these sentences significantly better than speakers of USE (yet, BrE and USE speakers assign fairly similar ratings). Once more, these sentences are rated best by speakers of HKE (interaction mediates effect by $B = 182$), followed by speakers of IndE ($B = 122$), which could be interpreted as general insecurities in language use resulting from an overall lower language proficiency.

Target sentences are rated highest in HKE, followed by IndE and SgE

For the target sentences with the converted forms, the ratings by all speakers are considerably higher than the ratings by USE speakers ($B = -613$ for USE). Nonetheless, even though statistically significant, the difference between BrE and USE speakers is again comparatively small ($B = 105$, $p < .001$). The fact that these two groups exhibit a similar rating behavior and rate the sentences containing converted nouns the lowest is in line with what can be expected (hypothesis 5, cf. section 8.4).

The rating behavior of the three groups of speakers of New Englishes differs significantly from that of the foregoing groups, yet it is not uniform. HKE speakers rate the target sentences the highest (compared to the control sentences, effect mediated by $B = 301, p < .001$), most likely due to the influence of the Chinese substratum. The effect is second highest, but already considerably lower, for speakers of IndE ($B = 206, p < .001$), closely followed by speakers of SgE ($B = 195, p < .001$).

Daily use of English predicts rating behavior

Whether participants use English on a daily basis or not influences their ratings. Participants' ratings of the control sentences or the sentences containing the 'foreign' non-standard features ('D') do not depend on their daily use of English. In contrast, the sentences containing the converted forms are rated significantly better by speakers who do not use English on a daily basis ($B = 67, p < .03$). It has to be noted that the group of speakers who stated that they do not use English every day mainly consists of HKE and IndE speakers. Consequently, it is likely that the results for this interaction are correlated with the results for the interaction between these varieties and the types of sentences. Yet, since the inclusion of daily use of English as a predictor yielded a model which was significantly better than the model not containing this predictor, daily use was kept in the final model.

Additionally, the sentences with the highly pervasive non-standard features ('A', $B = 73$) and the ones with the highly pervasive HKE and SgE features ('AAsian', $B = 108$) receive significantly better ratings. The effect is strongest for the sentences with the features pervasive in HKE and SgE ('AAsian', $p < .01$).

Differences between individual target sentences

As an analysis of the individual target sentences reveals (cf. figure E.7 in appendix E.8), the sentence containing *improve* is rated highest across all varieties. The sentence with *examine* is rated second-highest. Considering the complexity of acceptability judgment tasks and the many factors that can influence their outcome (such as the frequency of the words contained in the stimuli as well as their semantic content), it is at this point mere speculation whether these two lexemes are more successful than the other lexemes. Notwithstanding, *improve* and *examine* are also among the more successful converted forms in the corpus data.

Interpretation and discussion

USE participants disprefer conversion most

Speakers of USE rate sentences containing verb-to-noun conversion lowest, compared to all other sentences types and speakers of all other varieties. This is in line with the finding in section 6.3.2 that the blocking effect is strongest in USE. As the corpus study has shown, statistical preemption is a powerful effect in USE, and if the corresponding derived noun is easily retrievable, speakers of USE strongly disprefer the converted noun. Since the target sentences contain converted nouns that have been converted from base verbs of a comparatively high frequency, it is to be expected that USE speakers rate these sentences lowest of all groups.

Substrate transfer is moderated by degree of institutionalization

In line with what has been hypothesized above (hypothesis 1), speakers of the new varieties assigned the sentences containing converted nouns higher ratings than speakers of native varieties. Furthermore, as predicted (hypothesis 2, 3), the ratings provided by speakers of the new varieties show considerable cross-varietal differences, with speakers of HKE assigning the sentences containing V>N conversions much higher ratings than speakers of SgE and IndE. The finding that conversion of verbs to nouns is judged more acceptable by speakers from Hong Kong than by speakers from Singapore underlines the results of the corpus analysis. It seems that also in the judgment of acceptability, the effect of a substratum is moderated by the degree of institutionalization of English, prompting SgE speakers to assign lower ratings than HKE speakers, despite the similar contact ecology of Hong Kong and Singapore.

‘Usage despite awareness’ in Singapore

The finding that SgE participants rate the sentences containing features that are highly pervasive in SgE and HKE (‘AAsian’) comparatively low is not too surprising in light of the linguistic background of this group of participants in particular as well as the language policy adopted in Singapore more generally: Almost all participants indicated that they use English on a daily basis and have acquired it from birth, resulting in a (very) high language proficiency. Also, the fact that many participants have a high or very high level of education makes these participants prone to adapting a norm-conforming linguistic behavior. On a more general level the SgE participants could have a high awareness of these highly salient

features of SgE due to the *Speak Good English Movement* (cf. section 1.1.3), which is an important part of the language policy in Singapore and which seeks to approximate the local standard to an exonormative standard. In a stabilizing setting such as Singapore, the incoherence of corpus findings and rating results can be assumed to constitute an instance of ‘usage despite awareness’. That is, even though some features receive low ratings in acceptability judgment tasks, they are used systematically (in corpora). This difference between rating results and actual usage as evidenced in corpora shows that acceptability judgment causes raters to consciously access their metalinguistic knowledge, which is presumably influenced by language policy and by orientation towards an exonormative standard. In contrast, the systematicity of use of a feature in language production, particularly in near-spoken (web-) registers, is a reflection of spontaneous, unconscious decisions.

Lower English proficiency explains findings for IndE

While substrate influence could be assumed to lead SgE speakers to rate the sentences with V>N conversion comparatively higher than the control sentences (compared to the USE speakers), substrate influence is implausible as an explanation for the results obtained from the IndE speakers. Once more, the degree of institutionalization of English seems to play a key role in explaining the outcome of the rating. A lower degree of institutionalization leads to a lower proficiency in the English language, which in turn can culminate in speakers either having a lower language awareness and being insecure about what is ‘acceptable’ (hence the higher rating of the ‘D’ and ‘AAsian’ sentences even though these features do not occur at all or not to the same extent in IndE) or favoring conversion, a morphologically simple process also found in learner varieties.

That this effect is stronger in IndE than in SgE is probably due to the specific dataset. Even though SgE and IndE are often located at the same stage in the Dynamic Model, the participants from the two countries show markedly different language backgrounds: The number of IndE participants not using English on a daily basis is higher than that of SgE speakers who do not make daily use of English (cf. figure 8.3). Also, while almost all SgE participants give English as one of their native languages, the number of IndE participants with English as their native language is below 25% (cf. figure 8.2). It is therefore plausible to assume that the effect of (lower) proficiency is stronger in IndE than in SgE.

In addition, the corpus analysis has revealed that the blocking effect is less strong in IndE, leading to higher odds of occurrence of V>N conversion in IndE. This tendency is reflected in IndE speakers rating sentences containing converted forms a little higher than SgE speakers. Summarizing, the results obtained for IndE and SgE do not fully line up with hypothesis 4,

according to which differences between IndE and SgE are hypothesized to be mainly due to substrate influence. As the language background of the participants suggests, despite both varieties often being situated at stage 4 of the Dynamic Model, the role that English plays in the lives of the participants constituting the sample from both countries differs considerably.

Daily use of English in new varieties induces rating behavior similar to native varieties

Overall, SgE speakers' ratings are closest to the ratings provided by speakers of native varieties. Seeing that SgE is the most indigenized of the new varieties, this result is also indicative of a strong proficiency effect. The existence of this effect of proficiency is further corroborated by the interaction of the daily use of English and the type of sentence rated (last block in table 8.6). The target sentences receive significantly higher ratings by speakers who do not use English on a daily basis.

Moreover, a proficiency effect can also be noted by comparing the dispersion of the rating values across varieties (cf. figure 8.5). Whereas the native varieties present comparatively homogenous rating values (as evidenced by the comparatively small sizes of the boxes), the rating results by the participants of new varieties are much more dispersed e.g. for the control sentences or the 'D' sentences, where all groups were expected to assign a low rating. The higher dispersion points to a lower intra-group consistency, which could be indicative of more insecurities when it comes to judging language. This tendency is strongest for HKE, the variety with the lowest rate of English native speakers and the highest rate of speakers who do not use English every day.

Most sociolinguistic variables and background in linguistics are not significant predictors

A further point which is interesting to note is that apart from the participants' language background none of the other variables related to the participants' background significantly influences the rating result. Age, gender, education, and whether participants have a background in linguistics are not significant predictors. Also, the part of the variance that is accounted for by the individual participant is fairly small (roughly 11%).

Summary of the rating task

The rating task was designed to find out how speakers of various varieties judge the phenomenon of verb-to-noun conversion. It has been hypothesized that an increased exposure to V>N conversion will prompt a higher acceptability of the phenomenon. The results show that indeed in the nativizing setting that e.g. HKE represents, a higher systematicity as ev-

identified by a higher frequency in corpora comes with higher acceptability ratings. In the particular case of HKE this is most likely due to the influence of the substratum but also to some extent to the comparatively low degree of nativization of English in Hong Kong. The latter point becomes particularly obvious when comparing HKE to SgE, where V>N conversion is rated significantly lower, despite the fact that both varieties have emerged from comparable linguistic ecologies.

The degree of institutionalization is intricately linked to the level of language proficiency. Speakers of less institutionalized varieties show a lower level of English proficiency which in turn finds expression in the rating results, not only in the absolute rating values but also in the greater heterogeneity of the results. This proficiency effect reflects particularly strongly in IndE, for which the results deviate from what could be anticipated on the basis of the classification of IndE as an endonormatively stabilized variety (a classification frequently provided in the literature, cf. section 1.1.3). That the results for IndE differ from the expected can be attributed to, first, the fact that the stages of the Dynamic Model are too coarse and that SgE and IndE differ in how institutionalized they are, even if both might be subsumed under the heading of endonormatively stabilized varieties, and second, that the “convenience sample” of participants (Schnell et al. 2011: 377) is not representative of the Singaporean and Indian population at large.

As predicted, speakers of native varieties do not differ significantly in their rating behavior and find V>N conversion least acceptable, compared to (almost) all other types of sentences and to all other varieties. Whether these results can be confirmed by the maze task, which assesses processing, is the focus of the next section.

8.9.2 Maze task

The results of the maze task are presented in two blocks. The first is concerned with the reaction latencies for the converted vs. derived forms on the word level, while the second is an analysis of observations of what occurs at the sentence level.

Reacting to converted forms

For the linear regression model for the maze task, only the reaction times to the converted or derived forms in correctly assembled sentences were considered. The following are the predictor variables specific to the maze task:

typeStimulus (Deriv) Is the stimulus a converted (Conv) or a derived (Deriv) form?

logRTprev This numeric variable captures the logarithmically transformed reaction time to the word immediately preceding the converted or derived form. According to Baayen and Milin (2010: 19), the preceding latency influences the current latency in such a way that if the reaction time to the previous stimulus is high, the current reaction time is also going to be higher. Including this (supposedly significant) predictor leads to a model with a smaller residual error, “allow[ing] a more precise estimation of the contributions of the other, theoretically more interesting, predictors” (ibid.: 21). As will be shown below, this predictor did indeed turn out to be highly significant in the regression model ($B = 0.135, p < .001$). Including it yielded a significantly better model, as a comparison of the AIC and BIC showed.²³

random effect: previous, short: prev This predictor relates to the previous one. It captures whether the participant, in order to choose the stimulus, pressed the same or the other arrow key as for the word immediately preceding the current word. In the literature on RT experiments (cf. e.g. Crump et al. 2013: 5) an increase in reaction times that comes with performing a task different from the one which has just been performed is called task-switching cost. Choosing the other arrow key might induce such a cost. However, even though technically it is possible to record which key was pressed, there is no way of determining how the key was pressed, potentially introducing a number of confounding factors. Among these are, for example, the handedness of the participants²⁴ or the strength of the fingers with which participants pressed the keys. If they used the index finger of the right hand to press the left arrow key and the ring finger to press the right key, this might lead to different RTs simply due to the fact that for many people the index finger is better trained than the ring finger. Due to all these uncertainties related to the key press, the factor of pressing the same key again or switching the key was included as a random effect.

random effect: lexeme In both tasks, trends are expected to emerge independently of individual sentences which participants rate or build. Nonetheless, unlike the model for the rating results, the linear regression for the maze task is only fitted to the reaction times to the converted and derived forms. Therefore, lexeme (instead of sentence ID) is included as a random effect. Furthermore, RTs to individual words depend on various

²³The AIC for the model including the previous reaction time is 3119.27, the AIC for the model without the previous reaction time is 3174.28. The BIC is 3249.67 with and 3298.47 without the reaction time to the previous word.

²⁴The arrow keys are usually on the right-hand side of the keyboard, which might disadvantage left-handed participants.

frequency measures, e.g. those given in Baayen and Milin (2010: 16). The construction of adequate stimuli is not only a highly complex task, but can also be problematic when investigating varieties of English. Most probably the construction of artificial stimuli would come at the expense of accurately representing converted forms as they are found in varieties of English. For the present purposes it was therefore deemed appropriate to include only slightly modified sentences in the experiment, drawn from major corpora of the English language (see above), and, as a consequence, to include lexeme as a random effect in the regression model (ignoring potential effects of frequency on the result).

As recommended by Baayen (2008: 31), RTs were logarithmically transformed to obtain (roughly) normally distributed values. Furthermore, following Enochson and Culbertson (2015: 5–6) decisions with RTs below 100ms were discarded. This yielded the reaction times displayed in the histogram in figure 8.6.

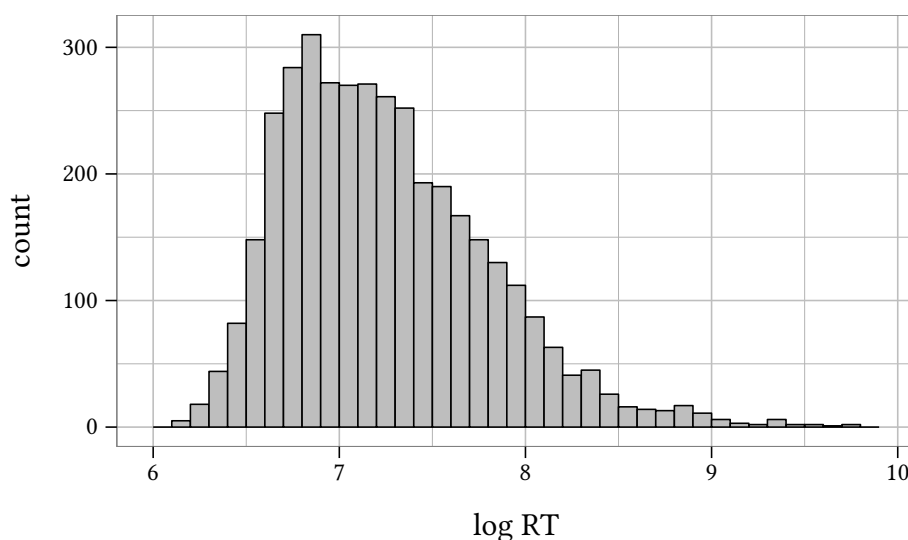


Figure 8.6: RTs in the maze task (binwidth = 0.1)

As with the rating task data, a model was fitted including all of the above-mentioned predictors. The formula is given in 8.8. Table 8.7 gives an idea of what the data input into the linear regression look like.

$$\begin{aligned}
 \log RT \sim & \text{variety} + \text{typeStimulus} + \log RT_{\text{prev}} \\
 & + \text{linguistics background} + \text{age} + \text{gender} + \text{education} \\
 & + \text{variety} : \text{typeStimulus} \\
 & + (1 \mid \text{Worker ID}) + (1 \mid \text{lexeme}) + (1 \mid \text{prev})
 \end{aligned}
 \tag{8.8}$$

Table 8.7: Subset of the coded data

WorkerID	lexeme	prev	logRT	typeStimulus	variety	logRTprev	linguist	age	gender	education	
1	55a654f5ac922a	government	same	6.6477	deriv	US	6.4216	no	4.97	female	bachelor's degree
2	55a654f5ac922a	expectation.	same	7.0246	deriv	US	6.4085	no	4.97	female	bachelor's degree
3	55a654f5ac922a	acceptances	other	6.9689	deriv	US	6.6682	no	4.97	female	bachelor's degree
4	55a654f5ac922a	approval	other	6.4265	deriv	US	7.2204	no	4.97	female	bachelor's degree
5	55a654f5ac922a	calculation	same	6.9698	deriv	US	7.2584	no	4.97	female	bachelor's degree
6	55a654f5ac922a	continuation	same	6.5367	deriv	US	6.6187	no	4.97	female	bachelor's degree
7	55a654f5ac922a	consumption	other	6.6333	deriv	US	6.5162	no	4.97	female	bachelor's degree
8	55a654f5ac922a	possession.	same	6.4739	deriv	US	6.5396	no	4.97	female	bachelor's degree
9	55a654f5ac922a	enhancement	other	7.5126	deriv	US	7.0121	no	4.97	female	bachelor's degree
10	55a654f5ac922a	provision	same	6.8701	deriv	US	6.4892	no	4.97	female	bachelor's degree
11	55a654f5ac922a	creation.	same	6.7979	deriv	US	6.4922	no	4.97	female	bachelor's degree
12	55a68bfa1f288a	discovers	same	6.9537	conv	US	7.3059	yes	1.97	female	master's degree or higher
13	55a68bfa1f288a	agree	other	7.5060	conv	US	6.5323	yes	1.97	female	master's degree or higher
14	55a68bfa1f288a	begin	other	7.2027	conv	US	6.4441	yes	1.97	female	master's degree or higher
15	55a68bfa1f288a	choose.	same	7.3330	conv	US	6.4599	yes	1.97	female	master's degree or higher
16	55a68bfa1f288a	examine	other	6.8341	conv	US	6.2841	yes	1.97	female	master's degree or higher
17	55a68bfa1f288a	facilitate	other	7.1452	conv	US	7.3159	yes	1.97	female	master's degree or higher
18	55a68bfa1f288a	improve	same	7.2145	conv	US	6.5381	yes	1.97	female	master's degree or higher
19	55a68bfa1f288a	intend	same	7.1229	conv	US	6.6946	yes	1.97	female	master's degree or higher
20	55a68bfa1f288a	invest,	other	6.6708	conv	US	6.6846	yes	1.97	female	master's degree or higher
21	55a68bfa1f288a	realize	other	6.7298	conv	US	6.8352	yes	1.97	female	master's degree or higher
22	55a68bfa1f288a	receive	same	7.1309	conv	US	6.5820	yes	1.97	female	master's degree or higher
23	55a68bfa1f288a	remind.	other	7.0397	conv	US	6.5103	yes	1.97	female	master's degree or higher
24	55a68bfa1f288a	suggestion	same	7.4319	deriv	US	6.7370	yes	1.97	female	master's degree or higher
25	55a68bfa1f288a	survival	same	6.7776	deriv	US	6.7044	yes	1.97	female	master's degree or higher
26	55a68bfa1f288a	threat	other	6.7890	deriv	US	6.7166	yes	1.97	female	master's degree or higher
27	55a68bfa1f288a	communication	other	6.3919	deriv	US	6.5681	yes	1.97	female	master's degree or higher
28	55a68bfa1f288a	government	other	6.6346	deriv	US	6.9660	yes	1.97	female	master's degree or higher
29	55a68bfa1f288a	expectation.	other	7.0724	deriv	US	6.4998	yes	1.97	female	master's degree or higher
30	55a68bfa1f288a	acceptances	same	7.0842	deriv	US	6.6107	yes	1.97	female	master's degree or higher

Yet, linguistics background turned out to be a non-significant predictor. Therefore, it was removed, yielding the model with the corresponding formula in 8.9.

$$\begin{aligned} \log\text{RT} \sim & \text{variety} + \text{typeStimulus} + \log\text{RTprev} \\ & + \text{age} + \text{gender} + \text{education} \\ & + \text{variety} : \text{typeStimulus} \\ & + (1 \mid \text{Worker ID}) + (1 \mid \text{lexeme}) + (1 \mid \text{prev}) \end{aligned} \tag{8.9}$$

Model criticism and outlier removal

Since reaction time data are highly susceptible to noise and influenced by a multitude of factors (cf. section 4.3.2), it is sensible to scrutinize this particular dataset and the model fitted to it for potential outliers. Generally, “outlier removal before model fitting is not necessary”, “if the precondition of normality [of the distribution of RTs] is well met”, as Baayen and Milin (2010: 16) state. As the RT data are roughly normally distributed (cf. figure 8.6), no outliers were removed. This resulted in a model with a fit that was not ideal. Residuals were larger than generally recommended, as can be seen in the residual plots in appendix E.9. Generally, outliers should not be removed if they are the result of variability in measurement.²⁵ In contrast, if they are experimental artifacts, removing them can be considered justified (cf. Osborne and Overbay 2004).²⁶ Particularly when it comes to reaction times, Baayen (2008: 257–258) and Baayen and Milin (2010: 26) recommend that all “datapoints [sic] with standardized residuals exceeding 2.5 standard deviations” (ibid.) be removed from the data input into the model. This trimming procedure resulted in the removal of 2.3% of all data points. This number is small, yet a small number of outliers can suffice to either generate an effect that is not “supported by the majority of data points” or “mask[...] an effect that is actually supported by the majority of the data points” (ibid.). The model fitted to the subset of the data, hereafter the ‘trimmed model’, showed a better fit with more normally distributed residuals (cf. residual plots in appendix E.9). A histogram of the remaining data points is plotted in figure E.9 in appendix E.11. Most coefficients have similar estimate sizes and remain significant or insignificant, however, some coefficients change in significance from the first model to the trimmed model. This will be addressed in the ensuing description of the results. Table E.6 in appendix E.10 gives the coefficients of the linear regression fitted to the original dataset, while table 8.8 presents the results of the model fitted to the trimmed dataset. The scaled residuals and the random effects for both models are reported in tables E.7 and E.8 in appendices E.10 and E.12, respectively.

²⁵Therefore, no outliers were removed from any of the other regression models.

²⁶The question of whether outliers should be removed and if so, how they should be identified, is highly complex, as Osborne and Overbay (2004) illustrate.

Figure 8.7 is a graphic representation of the results of the maze task. It is similar to figure 8.5. Logarithmically transformed reaction times are displayed separately for the two types of stimuli (derived forms on the left, converted forms on the right). For every variety, there is a separate box. The boxes again represent the interquartile range, the whiskers 1.5 times the interquartile range, the horizontal line the median, and the diamond the mean. Values that do not fall within 1.5 times the interquartile range above or below the third or first quartile are given as dots.

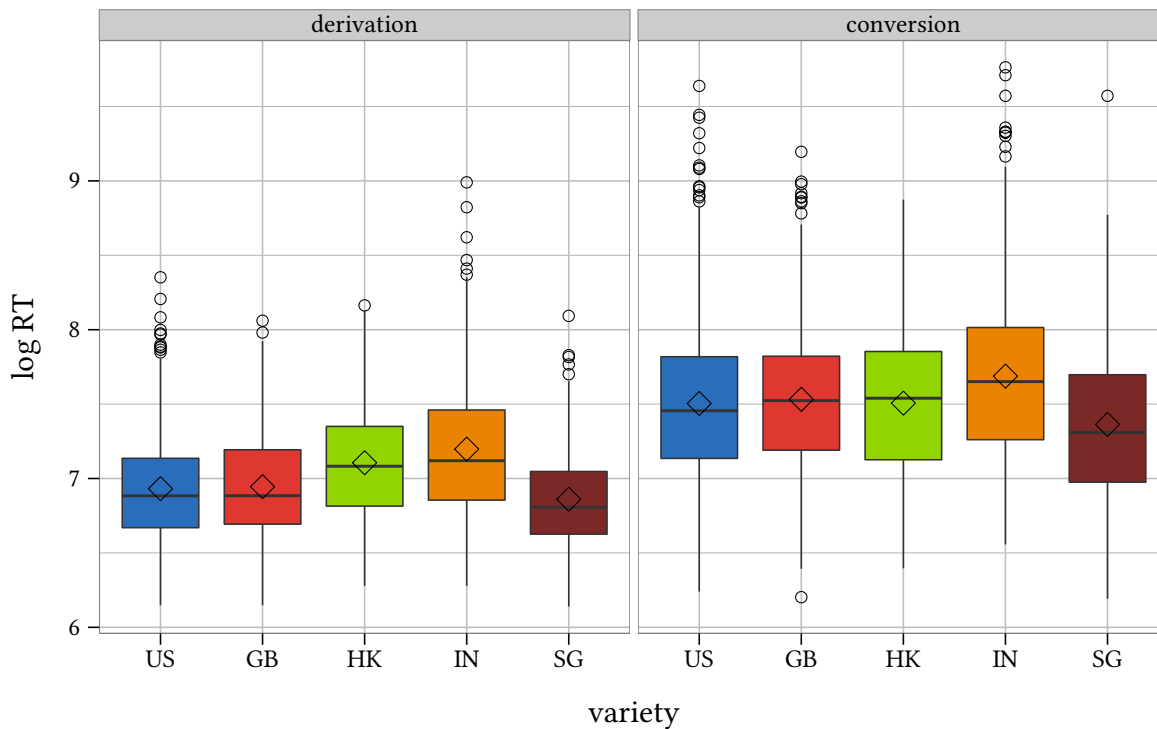


Figure 8.7: Boxplots of RTs by variety and type of stimulus. For each box, the horizontal line represents the median and the diamond represents the mean.

Results

Speakers of native varieties and of SgE react fastest to control stimuli

The intercept in this model refers to the reference variety, USE, and to the stimulus type of derived forms.²⁷ The block right below the intercept gives the estimates for the derived forms for the other varieties. As expected, speakers of BrE do not differ significantly from

²⁷Note that this value (given in logarithmically transformed reaction times) does not correspond to the mean RT to derived forms of USE speakers, but is the modified mean value. In other words, the intercept also encompasses the other predictors, i.e. represents the mean for female participants with a master's degree or higher, an age of 0 (recall that age is centered), and an average previous reaction time.

Table 8.8: Maze Task (trimmed dataset)

	Estimate	Std. Error	df	t val	p	
intercept						
(Intercept)	5.939	0.13	49.55	45.56	0.000	***
varieties						
varietyGB	0.053	0.04	190.72	1.21	0.230	
varietyHK	0.161	0.06	213.36	2.83	0.005	**
varietyIN	0.211	0.04	198.45	5.07	0.000	***
varietySG	-0.048	0.05	188.54	-1.04	0.302	
type of stimulus						
typeStimulusConv	0.533	0.05	69.96	11.21	0.000	***
metadata						
age	0.003	0.00	143.25	1.90	0.059	.
genderMale	0.067	0.03	146.91	2.25	0.026	*
genderOther	-0.081	0.17	138.35	-0.47	0.637	
educationBachelor's degree	0.088	0.03	148.39	2.80	0.006	**
educationHigh school diploma	-0.021	0.05	148.80	-0.44	0.657	
educationHigh school, no degree	-0.023	0.05	154.92	-0.43	0.671	
previous RT						
logRTprev	0.135	0.02	3658.03	8.11	0.000	***
variety : type of stimulus						
varietyGB : typeStimulusConv	0.002	0.03	3464.73	0.06	0.949	
varietyHK : typeStimulusConv	-0.104	0.04	3480.98	-2.33	0.020	*
varietyIN : typeStimulusConv	-0.082	0.03	3472.13	-2.62	0.009	**
varietySG : typeStimulusConv	-0.075	0.04	3470.88	-2.14	0.032	*

speakers of USE in how fast they respond to the derived forms, and neither do participants from Singapore ($p < .23$ for BrE, $p < .31$ for SgE).

Derived forms are processed more slowly by speakers of HKE and IndE

Nevertheless, for HKE and IndE the situation plays out differently. Participants of both varieties show significantly higher reaction times than speakers of the native varieties or of SgE, with IndE participants taking longest (intercept + 0.211, $p < .001$). For speakers of HKE the higher reaction times (HKE: intercept + 0.161, $p < .01$) might easily be attributed to the influence of the Chinese substratum, which is assumed to induce a dispreference for morpho-

logically more complex forms. However, as has previously been pointed out, the influence of the substrate cannot be the only reason for these reaction times, otherwise IndE speakers would be expected to react much faster.

Speakers of all varieties react more slowly to V>N conversion

The results further reveal that verb-to-noun conversion is processed considerably more slowly in all varieties ($B = 0.533, p < .001$). Yet, the strength of this effect varies significantly between varieties, as the last section of table 8.8 shows.

Speakers of New Englishes react faster to target stimuli containing conversion, with SgE coming closest to native varieties

The last block of table 8.8 gives the results for the reaction times to converted forms in BrE, HKE, IndE, and SgE compared to USE. As with the derived forms, BrE participants' reaction times do not deviate significantly from USE participants' ($p < .95$). By contrast, participants of the new varieties differ in that they react markedly faster to the converted forms. Nonetheless, similar to the derived forms, SgE participants are once again closest to participants from the native varieties. In the model fitted to the original dataset the difference in reaction times to converted forms between USE and SgE does not even reach significance ($p < .17$, cf. table E.6). After trimming the dataset, the effect becomes significant ($p < .04$), yet it is smaller than in any of the other new varieties ($B = -0.075$).

Fastest reaction times to conversion in HKE

In HKE, the difference in reaction times to the derived vs. the converted forms is smallest. In other words, HKE participants react fastest to V>N conversion, compared to derivation and compared to speakers of all other varieties. However, this does not mean that HKE participants show the absolute lowest RTs to the converted forms. The absolute size of the effect is calculated adding all relevant estimates. For HKE, this means adding up the intercept, the estimate for varietyHK, the estimate for typeStimulusConv, and the estimate for the interaction of varietyHK and typeStimulusConv: $5.939 + 0.161 + 0.533 - 0.104 = 6.529$. For USE, the absolute value is $5.939 + 0.533 = 6.472$. Thus, US participants still react faster to the target stimuli, nevertheless, their RTs to the target stimuli are considerably higher compared to their RTs to the derived forms. The absolute difference in RTs to the derived vs. the converted forms is 0.429 for HKE and 0.533 for USE. These values illustrate that HKE participants react comparatively faster to verb-to-noun conversion.

8 Experimental validation of corpus results

The effect that converted forms are processed faster in HKE is significant in both models ($B = -0.104, p < .03$), suggesting that it is a stable effect that is borne out by the vast majority of data points. Also, the size of the estimate is the largest for HKE compared to the other new varieties. This clearly suggests that the Chinese substratum influences (at least in part) the time it takes HKE participants to react to conversion, which dovetails with hypothesis 2.

High reaction times for IndE in the trimmed model

The situation for IndE is very different for the two models. In the model based on the original dataset, the effect that IndE participants react faster to converted forms than other groups is barely statistically significant ($p < .06$). However, the model fitted to the trimmed data reveals a much more significant effect for the converted stimuli ($p < .01$), as well as an increased estimate ($B_{original} = -0.067$ vs. $B_{trimmed} = -0.082$). This change in significance level indicates that this effect must have been masked in the first model by a small number of extreme data points.

Younger participants react slightly faster

As regards the background of the participants (heading ‘metadata’), the effects that are present are comparatively small. In the model fitted to the original dataset age is a significant predictor of reaction times in that older respondents take minimally longer to respond. Nonetheless, this effect is marginal ($B = 0.004, p < .05$). In the model fitted to the trimmed data, this predictor loses its significance almost completely ($p < .06$), indicating a rather unstable effect. The fact that younger participants react faster and that older participants have higher reaction times is in line with what has previously been reported (cf. e.g. Ratcliff et al. 2004: 286–287). Considering its small size and the circumstance that it is nearly non-significant, it is highly unlikely that this effect points to a higher acceptance of the converted form by the younger participants.

Females react faster

A further effect is that male participants show slightly higher reaction times than female participants ($B = 0.067, p < .03$).²⁸ This advantage of females over males contrasts with what is generally reported in the literature (cf. Kosinski 2013, see below for discussion).

Participants with a bachelor's degree react more slowly

As far as the highest level of education is concerned, participants with a bachelor's degree have a significantly higher reaction time than participants with a master's degree or higher or participants without a bachelor's degree ($B = 0.088, p < .01$).²⁹

Interpretation and discussion

Degree of institutionalization crucially shapes RTs to derivation and conversion

The results of the maze task once again highlight that the degree of institutionalization of English plays a key role in the emergence of new varieties. The speakers of the new variety with the highest degree of institutionalization, SgE, display reaction times which come closest to the RTs of native speakers of English. This holds both for derivation and for conversion. This is in line with previous findings that SgE is the most nativized variety of the new varieties investigated and is therefore most similar to the native varieties (cf. chapter 6). This result once again underlines the claim that English is becoming the native language of more and more Singaporeans and that the status of SgE as an ESL variety has to be reassessed in light of current trends.

In contrast, speakers of less advanced varieties, IndE and HKE, exhibit significantly different RTs, reacting comparatively more slowly to the derived forms but comparatively faster to the converted forms. This is attributed to the difference in proficiency level directly resulting from a lower degree of institutionalization of English in these speaker communities.

For the derived forms, HKE and IndE speakers have comparatively lower reaction times. As derivation is a morphologically complex process, this effect might be due to a lower level of English proficiency of HKE and IndE speakers. As for example Pavesi (1998: 220–221,

²⁸The one participant who reported a gender other than female or male did not differ significantly from the female group. In regression modeling, it is common practice to conflate factor levels that occur only sparsely with other levels of the same factor (cf. e.g. Harrell 2015: 164). However, this seemed inappropriate in the case of gender, so that 'other' was kept as a separate factor level.

²⁹The groups of participants with a high school diploma or without it were considered separately in the model. Seeing that the number of individuals in both groups is comparatively small (cf. table E.4 in appendix E.6), a model where these two groups were combined was calculated. It did not yield significantly different results from the models in tables E.6 and 8.8 and is therefore not reported.

226) points out, morphologically more complex forms are only produced at a later stage in (second) language learning and simple forms (the products of conversion among them) are preferred at earlier stages. This could explain why the speakers with a lower proficiency in English need more time to react to the derived forms but less time to react to the converted forms.

Notwithstanding, for both HKE and IndE, substrate influence cannot be ruled out completely as an explanatory factor. As has previously been mentioned, Chinese does not make use of derivation frequently, which could be the reason for HKE speakers' unfamiliarity with the process. For IndE, Sailaja (2009: 76, 80) notes that "Indian languages are rich in compounding", which is why compounding might be preferred over affixation in IndE, causing longer latencies when it comes to coping with derived forms.³⁰

As far as the reaction times to the stimuli involving conversion are concerned, the effect of proficiency is particularly obvious for IndE, where substrate influence as an explanation for the observed effects seems unlikely.³¹ What is more likely, also with regard to the result of the rating task, is that the circumstance that IndE speakers react faster to V>N conversion is a consequence of their comparatively lower English proficiency (keeping in mind the above-mentioned peculiarities of this specific sample of IndE speakers).

Institutionalization moderates substrate transfer

The results of the maze task further corroborate the finding that the developmental stage of a variety has a greater power than substrate transfer in predicting the success of V>N conversion in general, and the reaction times to conversion versus derivation in particular. Once more, the results for HKE and SgE differ drastically, despite the fact that both varieties have emerged from comparable contact situations. Consequently, the influence of the Chinese substratum must be hypothesized to diminish with progressing nativization. This is evident when considering that the reaction times to both conversion and derivation are similar for speakers of SgE and the native varieties, but most dissimilar for speakers of HKE and the native varieties.

³⁰The great majority of the stimuli are indeed formed by suffixation so that this assumption is not unreasonable, although it might not hold for all stimuli (one exception is, for example, *choice*).

³¹Even though it cannot be ruled out in light of Sailaja's (2009: 76, 80) findings as well as Kachru's (2006: 115) note that V>N conversion does exist, at least to a minimal extent, in Hindi, the L1 of a substantial number of the Indian participants (cf. section 2.2.3).

Frequency influences language processing

In all varieties, the converted nouns are processed considerably more slowly than the derived forms. This result is not surprising keeping in mind that across varieties derivation is by far the more common of the two processes. In line with usage-based argumentation, the more frequent process is predicted to be better entrenched in speakers' minds. Therefore, this process is also more easily accessible, resulting in lower reaction times. This is what is observed for derivation. By contrast, V>N conversion is the less common of the two processes and occurs with a much lower frequency (cf. chapter 6). Hence, this process is predicted to be less well entrenched, making it less accessible. This claim is substantiated by longer reaction times to V>N conversion.

Inconclusive evidence for the influence of gender and education on RTs

While the result for the influence of age on the reaction times tallies with the literature, the effects found for gender and education run contrary to effects generally reported.

Usually, male participants have been noted to show faster reaction times (cf. Kosinski 2013). Nonetheless, it seems that the evidence on sex differences is not conclusive. Walentin (2009: 181), in a review of studies concerned with "sex differences in verbal abilities and language cortex" (ibid.: 175), summarizes that overall there is no statistically significant difference in language function or ability in healthy adults. If differences exist, these are rather to be sought in different processing or learning strategies. Therefore, the present effect should preferably be attributed to the idiosyncrasies of this particular sample. The distribution of female and male participants per variety also points in this direction. The IndE group, the slowest of all, consists primarily of male speakers (> 70%). In contrast, the fastest groups, SgE, BrE and USE, show a higher proportion of female participants (70% on average). The effect of gender must thus be assumed to be partly confounded with the effect of variety.

As regards the effect of education on reaction times, participants with a bachelor's degree are the only group to show significantly slower RTs. For the group of participants without a bachelor's degree, it is plausible to presume that it overlaps to a large extent with the younger participants. A Spearman rank correlation test of education and age shows that a significant correlation of medium strength holds between these two predictors, with younger people having a lower level of education ($\rho = 0.399, p < .001$). It would therefore not be surprising if this group reacted faster than the slightly older participants who have already obtained a bachelor's or master's degree. Indeed, the estimate points in this direction, even though the

difference in RTs between participants with a master's degree and this group does not reach statistical significance.

That the participants with a bachelor's degree react significantly more slowly than participants with a master's degree cannot be explained by the age of the participants, considering that the latter are older than the former. However, it could be that the participants who have obtained a master's degree or higher are faster at processing in general or possess a higher ability to recognize repeatedly occurring patterns, since they learn faster. After having been presented with a certain number of stimuli, participants with a master's degree may have recognized the aim of the maze task and may have responded more quickly upon seeing the same pattern (verb used as noun) again. What must not be neglected as another potential explanation is the fact that due to the recruiting process the sample of participants is highly unsystematic, so that the effect of education might be due to mere chance.

Reacting to sentences containing converted nouns

Some sentences more difficult than others

An analysis of the answers to the individual sentences reveals similarities as well as differences across varieties. All groups of participants had a rate of correct responses of lower than 50% for sentence 3. This is supposedly due to the ambiguous nature of the N+N compound *vacation offers*, which must often have been interpreted as part of the subject and the predicate. Also, sentence 2 was constructed without error by less than 50% of the participants from the US, UK and India. It is highly likely that once again the N+N compound (*Falklands people*) caused a considerable amount of confusion among the participants. For a potential follow-up study, it is therefore recommendable to avoid N+N compounds or to present them in the same step, as was done in sentence 7 (*'customer-centricity'*), for example.³²

For the participants from Hong Kong, sentences 7 and 10 proved difficult. For sentence 7, the challenge could lie in that it is the second longest sentence (20 steps). Sentence 10 is the only sentence in the set of stimuli which consists of a pseudo-cleft construction. As Winkle (2015) points out, this construction is comparatively infrequent in HKE, which could explain why participants had difficulties building this sentence.

IndE participants were particularly challenged by sentence 20. Two potential reasons could be that the sentence contains a subordinate clause as well as a coordinated predicate (*can and can't do*). Of these, the second seems more likely considering that in general IndE

³²Also cf. Forster (2010: 353–354) for the potential processing difficulties of structural ambiguities.

participants succeeded in building other sentences with subordinate clauses. A chart with the rate of correct responses to each sentence (figure E.10) can be found in appendix E.13.

Highest error rates in HKE and IndE

The error rates per variety are displayed in table 8.9. What these show is that participants speaking HKE or IndE have a much lower rate of correctly built sentences than participants speaking native varieties or SgE. The fact that speakers of HKE and IndE not only react more slowly, but also perform worse is another finding in support of the above-mentioned proficiency cline between speakers of native varieties and SgE on the one hand and speakers of HKE and IndE on the other hand. This cline in proficiency is a direct result of whether speakers use English every day or not, which is in turn directly linked to the degree of institutionalization of English in the respective region or country.

Table 8.9: Rate of correct responses per variety

variety	percentage of correct sentences
USE	80.2%
BrE	71.8%
HKE	62.5%
IndE	62.9%
SgE	81.4%

Cross-varietal differences in sentence processing patterns

This difference in processing between speakers of well-established and less well-established varieties is not only found for speed and accuracy but also extends to sentence processing patterns more generally, as figures E.11 through E.14 in appendix E.13 show. They give the reaction times for every word in the sentence, once in the control condition and once in the condition with the converted form. These plots represent a selection of the sentences used in the maze task and are analyzed in a more qualitative way in the following.

Figures E.11 and E.14 illustrate how the difference in RTs to converted and derived forms is smallest (going down to almost non-existent) in HKE compared to any of the other varieties. Furthermore, the other varieties exhibit a much higher variance in RTs (as indicated by the errorbars representing one standard error above and below the mean) to the converted stimulus than HKE. These patterns support the assumption that there is considerable

influence from the Chinese substratum in HKE, which is not present in the more nativized Chinese-substratum variety SgE.

Figure E.12 confirms the hypothesis that speakers of the native varieties BrE and USE have most difficulty in processing the converted forms, prompting longer RTs, compared to the new varieties. From a usage-based perspective this finding is not surprising considering that V>N conversion is least frequent in the native varieties (cf. chapter 6). More experience with a linguistic phenomenon, that is a higher frequency of occurrence of it, results in automatization and subsequent faster retrieval. As speakers of BrE and USE have less experience with V>N conversion than speakers of HKE, SgE, and IndE, their reaction times are lower.³³

For HKE, both the converted as well as the derived form in the sentence visualized in figure E.12 elicit high absolute RTs. This is not the case for the other sentences displayed in appendix E.13. It could be that this effect is due to the circumstance that the near-synonym *receipt* is not formed by suffixation but by ablaut, analogous to *choose* > *choice*. This form, as it is non-compositional, might be processed differently, and therefore potentially pose more problems for less proficient speakers. That this tendency is also visible, albeit to a smaller extent, in IndE, the second least established variety, also points in this direction.

A further point worth noting about this sentence is that HKE speakers show the highest mean RT to the word *claiming* across all varieties. While this might be due to the “convenience sample” of participants or the lexeme occurring with a lower frequency in HKE, it might also be symptomatic of the processing burden that the present participle imposes on speakers of HKE. If this were the case, it would once again support the claim that HKE is heavily influenced by the non-inflecting Chinese substratum but also by the fact that it is a variety that is not entirely nativized yet. However, making this claim on the basis of just one data point is, of course, highly speculative.

Methodological considerations

The sentence represented in figure E.13 serves to address some methodological aspects. The first issue is the quality of the data. Particularly for the sample of IndE speakers there is little consistency in how this sentence is processed. The curves for IndE speakers do not overlap as accurately as they do for the USE speakers. Since the differences in RTs for any of the words except for the last (which is the target or control stimulus) should, in theory, be identical, it can be assumed that the differences observed are the result of inter-speaker variation.³⁴

³³Since the rest of the sentence is processed almost identically by all speakers, regardless of whether they are assembling the sentence containing the derived stimulus or the converted form, this finding is presumably not an artifact of the sampling procedure.

³⁴The same is supposedly the case for the lexeme *outcome* in HKE.

In order for this variation not to bias the regression model, it is indispensable to sample a high number of participants (cf. Baayen 2008: 159). The second issue is that of constructing adequate stimuli for a reaction time task. In this sentence, speakers of new varieties show higher absolute RTs to the lexeme *possibly*. As a search in GloWbE reveals, *possibly* on its own as well as premodified by *very* occurs less often in these varieties. This illustrates that processing and reaction times are highly susceptible to effects of frequency, which makes constructing stimuli that are processed identically across five varieties of English as distinct as the ones investigated here a highly complex if not unfeasible task. (Even more so if the stimuli are to represent speech typical of a new variety.) In order to avoid that the frequencies of individual lexemes influence the outcome of a statistical model in which frequencies of individual lexemes are not the object of investigation, it is preferable to include lexeme as a random effect, as is done here.

Summary of the maze task

The aim of the maze task was to show that speakers of different varieties process verb-to-noun conversion differently. The results suggest that this is indeed the case. In line with the results of the rating task and with the findings on conversion from previous chapters, it has been found that the reaction times of USE and BrE speakers do not differ significantly (cf. hypothesis 5). Also, speakers of the native varieties react significantly more slowly to V>N conversion, reflecting their higher response uncertainty due to the reduced exposure to this construction (cf. hypothesis 1).

As in the rating task, HKE constitutes a special case in the group of varieties analyzed, with converted forms being reacted to much faster than in other varieties (cf. hypothesis 2). It is plausible to hypothesize that this is the result of transfer from the Chinese substratum. At the same time, it seems that the degree of institutionalization interferes with substrate influence. This interaction could explain why speakers of SgE react to V>N conversion much more slowly than speakers of HKE (cf. hypothesis 3).

For IndE, the results reveal highly significantly faster reaction times to the converted stimuli than in the native varieties. This is surprising in light of the fact that IndE does not have a Chinese substratum. However, it tallies with the findings from the corpus study in that the reaction times match the higher odds of occurrence of V>N conversion in IndE. The reasons for these results have to lie mainly outside the substrata, potentially in the level of proficiency of the participants as well as in the more liberal constraints that apply to word formation in IndE in general (cf. section 7.3).

8.10 Discussion and summary

The purpose of the experiment was to assess in how far the systematicity of the use of V>N conversion as witnessed in the GloWbE corpus translates to higher acceptability and faster processing of V>N conversion in individuals speaking HKE, SgE, IndE, BrE, and USE. The results of both tasks underscore the findings from quantitative and qualitative analyses of the corpus data presented in previous chapters.

The effect of frequency in both acceptability rating and processing is immediately apparent. In those varieties in which V>N conversion occurs more often, speakers are more accepting of the innovation and also respond to it faster. Yet, it has to be noted that because acceptability rating tasks inevitably access a speaker's metalinguistic knowledge, the results of the rating task are incongruent for SgE when compared to the results of the corpus analysis. This finding has to be interpreted as an instance of 'usage despite awareness', that is, as a direct result of the tensions in emergent varieties between a conscious orientation towards an outside standard and growing indigenization leading to an endonormative standard (without speakers consciously taking note of it).

For HKE, language contact and the potential influence of the Chinese substratum are readily drawn on as an explanation for the findings. Comparing HKE to SgE adds the dimension of the socio-institutional status of English to the evolution of New Englishes. On the basis of the results presented in this chapter, the degree to which English is institutionalized in a particular region or society can be assumed to be crucial to the development of varieties. Even though HKE and SgE have emerged from comparable linguistic ecologies, V>N conversion (among other lexico-grammatical phenomena) yields markedly different ratings of acceptability and speed of processing in both varieties, which cannot be attributed to the influence of the Chinese substratum alone.

Adding IndE, a variety with non-Chinese, mostly synthetic substrata, to the picture as a basis of comparison challenges a largely contact-based approach even further. The results demonstrate that performance in both tasks depends to a large extent on the English proficiency of the participants, which is, in turn, a direct consequence of the role which English plays in the participants' daily life. From the data on their linguistic background it follows that the IndE participants use English considerably less often than SgE speakers and start learning the language at a later stage—a fact which directly reflects in the result that IndE speakers rate V>N conversion and stimuli containing features that do not occur in IndE ('D', 'AAsian') better than SgE speakers and also react to V>N conversion slightly faster.

The results thus reveal that the emergence of new varieties of English from language contact settings requires a multifactorial explanation, and that reverting to the substrate as the main explanatory factor is too simplistic. As Percillier (2016: 193) notes, “the fact that a given feature is plausibly explainable by substrate influence does not mean it is in fact the result of L1 transfer” (also cf. Mesthrie 2008: 634; Thomason and Kaufman 1988: 212–213). It is therefore advisable to entertain a skeptical approach towards language contact as the main explanatory factor, particularly as the present study demonstrates (both in the corpus-linguistic and the experimental part) that the effect that language proficiency—as a result of the degree of institutionalization—has ranks above substrate transfer in explanatory power. Rather than drawing on transfer from the substrate as “the default assumption”, it might be wiser to understand it as only one out of a group of various explanatory factors (Mesthrie 2008: 634).

As far as the methodology presented here is concerned, the present chapter has shown that web-based experiments can yield plausible results, even if the selection of participants is less careful than usual in psycholinguistic experimentation. As regards research in the field of World Englishes, web-based experimentation has proven to be a method that is both reliable as well as feasible in terms of time and financial resources. The results of the rating task and the maze task demonstrate that McGraw et al. (2000: 505) are indeed right when they claim that “[n]umbers [...] swamp noise”.

In summary, the experimental study has illustrated how verb-to-noun conversion is judged and processed differently in varieties of English. Trends revealed by the corpus analyses in previous chapters have been confirmed by the experimental tasks. The aim of the next chapter is to discuss and summarize the main findings from all chapters.

9 Discussion and conclusion

After presenting various analyses of verb-to-noun conversion, in this chapter I aim to connect the dots and discuss the findings from previous chapters with a view to the research questions laid out in section 1.2. First, the development of *DISCONNECT*, an exemplary case of verb-to-noun conversion in USE, is recapitulated, focussing on the constraints on conversion as well as semantic, stylistic, and formal aspects of this specific lexical item's development as a noun. Second, these same aspects are reviewed for verb-to-noun conversion in the Asian varieties analyzed, particularly pointing out differences between native and new varieties of English. Third, the question of the interaction between substrate influence and the degree of institutionalization of a variety is addressed. As both the corpus-linguistic and experimental studies have shown, the interplay between these two factors is the key to understanding verb-to-noun conversion in Asian varieties of English. Fourth, the suitability of the term *ESL variety* for HKE, SgE, and IndE is revisited with regard to the findings on verb-to-noun conversion. As has already been pointed out, the three Asian varieties studied present markedly different usage patterns of verb-to-noun conversion, resulting from their different socio-institutional settings. While HKE is a variety that is in parts similar to a learner variety, SgE is more native-like. Fifth, the way in which verb-to-noun conversion is processed is summarized, distinguishing between the speaker's and the hearer's perspective. Sixth, the methodology used in this study is evaluated, with an eye on the benefits of combining corpus-linguistic and experimental methods. A conclusion highlighting the contributions of this study to the fields of World Englishes research and research within the usage-based paradigm rounds off this chapter.

9.1 V>N conversion in USE

As the analysis of *DISCONNECT* and other select cases (cf. chapter 5) has shown, verb-to-noun conversion is only moderately productive in USE. *DISCONNECT* is certainly an exception, considering how far advanced the conversion process is for this verb (e.g. full nominal paradigm, diverse syntactic functions). It can be assumed that in this particular case, various factors

such as the infrequency of the verbal form and also the infrequency of the deverbal noun contribute to the success of the converted form. The other cases, *DIVIDE*, *INVITE*, and *PAY*, are similar to the case of *DISCONNECT*, but none of these converted nouns is equally successful.

What the analysis of *DISCONNECT* (cf. section 5.1) has also revealed is that verb-to-noun conversion is by no means unconstrained in USE. A usage-based approach has proven useful in the endeavor of shedding light on the constraints operating in the application of V>N conversion. The usage frequencies of the following two forms play an especially important role in promoting or constraining conversion.

First, the token frequency of the verbal base is crucial. The higher the frequency of the verb is, the less likely conversion to a noun becomes. This phenomenon has been labelled the conserving effect of frequency (cf. Bybee 2010: 24–25): When the ratio of verb to converted noun is highly skewed towards the former, conversion is dispreferred or even non-existent due to the fact that the base form will be highly entrenched as a verb. Associating this form with a different word class will come at a very high processing cost (cf. Ungerer 2002: 560–563). In these cases, the use of another nominal form, mostly the derived alternative, is more likely. An example of this scenario is *CONNECT*.

Second, the effect of verb frequency is overridden by the power of the effect of the corresponding near-synonymous derived noun (e.g. *disconnection/s*). As the examples for USE have shown, the blocking effect, or statistical preemption, constrains verb-to-noun conversion even more strongly. The more frequent the near-synonym is, the less likely it is that the verbal base is converted to a noun, due to the fact that synonyms are generally avoided for reasons of linguistic economy. For *DISCONNECT*, the blocking effect was found to be relatively weak, whereas nominal *connect/s* is strongly blocked by the corresponding near-synonym *connection/s*.

The principle of the avoidance of synonyms further drives the semantic as well as the stylistic differentiation of the derived and the converted form in cases where the latter has established itself alongside the former (cf. sections 5.1.2 and 5.1.5). Shortly after its rise in frequency, *DISCONNECT*, the converted form, is predominantly used with the less technical, metaphorical meaning. As this particular case further illustrates, the derived and the converted form quickly evolve to occupy different registers, with the derived form being restricted to the more formal, academic register and the converted form occurring in more informal registers such as speech and newspapers.¹ This differentiation of (near-)synonyms is not unique to converted forms but well-attested for synonyms in general (cf. e.g. Leisi and

¹Nonetheless, their distribution can hardly be called complementary, seeing that *disconnect* (N) occurs more frequently than *disconnection*, even in the academic register.

Mair 2008: 46–48), which leads to the assumption that what can be shown for DISCONNECT is likely to apply to other instances of verb-to-noun conversion as well. Indeed, a differentiation between the converted and the derived form is also observed for e.g. DIVIDE. However, as this particular case reveals, it is not always the converted form which occurs more frequently in the informal registers. Nominal *divide* has evolved in such a way that this form is now most frequent in the academic register.

Another mechanism which has proven to be of high relevance to the success of conversion is the embedding of newly converted forms in frequently used constructions, often called chunks (cf. section 5.1.3). These units of language are cognitively well entrenched and therefore easily retrievable, which in turn leads to them being used with an even higher frequency. In the case of DISCONNECT, the co-occurrence of the converted form with the preposition *between* as well as its occurrence in the EXISTENTIAL construction have turned out to be of particular relevance for the spread of the novel form.

Finally, the corpus analysis has demonstrated that with the spread of a converted noun comes an extension in form, i.e. the elaboration of a full nominal inflectional paradigm (cf. section 5.1.4), and also an extension of usage contexts (cf. section 5.1.3). Both developments indicate a separation of the newly converted form from constructions with which it has frequently occurred before. Through an increase in frequency, the converted noun becomes more easily accessible as a construction in its own right and does not have to be embedded in larger constructions any more. This consequently leads to a less restricted usage of the construction, as has been pointed out for the particular case of DISCONNECT.

The case of DISCONNECT is prototypical and illustrates all facets of the verb-to-noun conversion process in an ideal way. Yet, the scenario that has been described for DISCONNECT (increase in usage frequency, emergence of a plural form, first restriction to certain constructions and contexts, then growing independence, semantic extension, etc.) is not obligatory. On the contrary, the conversion process does usually not proceed at such a quick pace as in this case, where the picture has changed drastically within the time frame of a mere twenty years.

The other case studies illustrate that for other verbs only parts of the above-mentioned scenario may happen. The case of INVITE, for example, shows that full conversion is possible, even though the converted form remains marginalized. It also shows that converted forms can be restricted to specific registers, in this case the magazine and newspaper register, potentially out of a preference for shorter or more innovative forms in these registers. The case of PAY demonstrates that it is not only the token frequencies of the verb or a potential blocking form that are decisive, but that the occurrence of a verb in many lexicalized compounds

can also restrict the availability of that verb for conversion. In summary, verb-to-noun conversion is a multi-faceted process, and it is only in conditions where all factors are favorable that it will proceed as successfully as in the case of DISCONNECT.

9.2 V>N conversion in Asian varieties

The subsequent corpus studies, both quantitative and qualitative, on native as well as Asian varieties (cf. chapters 6 and 7) have revealed that verb-to-noun conversion is a highly complex process that plays out differently in different varieties of English. Non-native varieties can be shown to be subject to some of the same factors at work in natively spoken varieties, but they also display differences, the causes of which are likely to be found in the specific contact situations and/or degrees of institutionalization.

The factor that seems to be operative in a similar way in all varieties is the blocking constraint (cf. sections 6.3.1 and 6.3.2). In all varieties, regardless of substratum or degree of institutionalization, the token frequency of the near-synonymous deverbal noun constrains the odds of conversion. The more frequent the competing form is, the less likely conversion becomes. Nevertheless, the effect size of the blocking constraint differs across varieties. Most notably, it seems to apply less in the case of IndE. In accordance with the literature on IndE word formation, the fact that conversion is comparatively unconstrained is attributed to a general tendency toward a more liberal use of diverse word-formation processes in this variety.

A further similarity between all varieties, whether native or non-native, is that the blocking constraint ranks above any effect the token frequencies of verbs might have on the odds of conversion. The strength of the blocking effect is also the reason why conversion is as elusive as it is. Even though substrate influence might lead to an increased probability of conversion, its influence is greatly diminished by the blocking effect of the deverbal noun.

Nevertheless, an effect for verb frequency was found; yet the size of this effect varies greatly between varieties, the notable exception in this case being HKE (cf. sections 6.3.1 and 6.3.2). The effect of the token frequency of the verb is significantly higher in this variety than in the other varieties. As has been illustrated in chapter 6, the less frequent a verb is in HKE, the more the odds increase that this verb will be converted to a noun.

These findings thus confirm that approaching the phenomenon of conversion, both in native as well as in non-native varieties, from a usage-based perspective is a worthwhile undertaking. Effects of frequency of various types do exist in all varieties, some applying to all of them, some limited to individual varieties only.

Further similarities between verb-to-noun conversion in USE and Asian varieties exist with regard to the effect of register (cf. section 7.6). As the cases of *DISCONNECT* and *DIVIDE* suggest, verb-to-noun conversion is a process that originates in the informal registers of the spoken mode. This is indeed confirmed by the data from IndE and SgE. In HKE, conversion is also present in formal contexts, which is ascribed to its developmental status (see below). The nature of the GloWbE corpus does not allow for a more detailed study of the registers in which verb-to-noun conversion occurs in new varieties, so that the question of the distribution of conversion across registers remains to be answered on the basis of other, more suitable evidence.

Also worth pointing out is the fact that many of the converted forms observed in the new varieties have not evolved as far as *disconnect* (N), particularly as regards semantics (cf. section 7.4.1). While the metaphorically extended meaning of *disconnect* (e.g. *disconnect between dream and reality*, COCA-ACAD, 2009) has become prevalent in recent years, lexicalization hardly occurs in the cases observed in the new varieties. The main meaning of the verb-to-noun construction seems to be that of reification, that is, of reconceptualizing verbs as nouns. This could, at least to some extent, be due to the fact that in the new varieties non-lexicalized conversions can take the place of gerunds or verbal nouns (cf. section 7.5).

A further aspect in which innovative conversions lag behind *disconnect* concerns form, namely the existence of a full nominal inflectional paradigm. For various of the innovative conversions, corresponding plural forms are not attested in the samples drawn from GloWbE. However, this observation is not based on representative evidence (if the singular forms of converted nouns occur infrequently, the plural forms are extremely rare) and is therefore necessarily speculative at this point. Another difference as regards form, highlighted in the qualitative analysis, is the dispreference of more complex constructions such as the gerund in verb complementation (e.g. *you can start bookmark those sites you like*, GloWbE-SG) or derivation in word formation in new varieties (e.g. *the purpose of our examine was to check...*, GloWbE-HK). This tendency presumably also triggers the higher preference for verb-to-noun conversion in these varieties (cf. section 7.5). The preference for morphologically less complex forms could be due to substrate influence (mostly for HKE and SgE) or to a more general simplification strategy in second-language varieties.²

Not unlike what has been shown for *disconnect* (N), V>N conversions in new varieties mainly occur in particular constructions, with the notable exception of IndE, where formal constraints seem to apply to a lesser extent (cf. sections 7.1, 7.2, 7.3). As the qualitative

²For a more detailed account of simplification tendencies in Asian varieties of English, cf. e.g. Seoane and Suárez-Gómez (2013: 11) for the perfective aspect and Terassa (in preparation) for past tense marking and plural marking.

analysis has revealed, V>N conversions are often embedded in the NOUN PHRASE construction. The preference for this construction can be attributed to the fact that it helps interpret the reconceptualization of a verb as a noun. The less ambiguous the constructional context is, the more easily a construction can be coerced. Contrary to what has been suggested in prior work (cf. Bernaisch 2015: 170–193; Hoffmann et al. 2011), the LIGHT VERB construction (*Mary had a walk in the garden.*) does not notably attract the V>N constructions analyzed in this study (cf. section 7.4.3). One reason could be that the meaning of an LVC does not overlap with the base verb to the same extent that the nominalization by conversion does. Teasing apart the details of the semantico-pragmatic meaning of V>N conversion compared to the LVC must at this point remain a topic for future research.

Similar to what the analysis of various cases of V>N conversion in USE indicates, V>N conversion in Asian Englishes is a gradual process with many facets. It is only in very rare cases that V>N conversion is so successful as in the example of *disconnect* (N), as the analysis of various verbs in new varieties corroborates. There are some verbs, like *examine*, which appear to be converted more often, sometimes even adopting additional meanings (as in e.g. *the scientific examine*, where *examine* means ‘study’, cf. section 7.4.1), while there are others, such as *allow*, that do not occur in the converted form at all (except for one instance). The reasons for the success of conversion could be manifold. In the morphological domain, the number of syllables or the derivational suffix required to form the near-synonymous noun could play a role. Also, phonetic constraints could influence the success of the process. Further, the semantic aspects of the base verb (e.g. concreteness of the concept) could be of importance. However, none of these aspects seemed to yield conclusive answers in the case of the twenty verbs under scrutiny here. A more systematic analysis of a larger database of verbs would be worthwhile to get a better idea of further constraints on V>N conversion.

9.3 The interplay of substrate influence and indigenization

In line with what has been hypothesized at the outset, the present study adds to the body of evidence which shows that dominantly contact-based accounts of the emergence of varieties are problematic. Undoubtedly, both the quantitative as well as the qualitative analysis of V>N conversion in non-native varieties have confirmed that the influence of the analytic Chinese substratum is one relevant factor in shaping the usage patterns in HKE and SgE. In both varieties, conversion is more likely than in the native varieties BrE and USE. Nevertheless, the numeric difference in the odds of conversion in HKE and SgE has proven to be highly significant, with SgE exhibiting comparatively lower odds of conversion. The same

has been found for the experimental tasks, in which participants from Hong Kong and Singapore differed in how acceptable they found V>N conversion and in how fast they reacted to it. A mere contact-based explanation, leaning heavily on the influence of the substrata to explain contact language grammar, turns out to be too simplistic to predict the profiles for V>N conversion. The assumption that the substratum contributes grammatical structure and the superstratum moderates the frequency of occurrence of that structure (cf. Bao 2005, 2009, 2010a,b) thus cannot be accepted as is, but has to be modified and expanded to accommodate the findings.

The alternative explanation that is readily at hand is the degree of institutionalization as operationalized by the stages of the Dynamic Model (cf. Schneider 2007). According to the prediction following from the Dynamic Model, varieties are expected to differ depending on their developmental stage. Indeed, the results of the corpus analysis and the experiment dovetail with this prediction: The native varieties show the lowest odds of V>N conversion, the lowest acceptability ratings and the highest reaction times. In contrast, HKE exhibits the highest odds and the highest degree of acceptability of conversion and the lowest reaction times. SgE and IndE cover the middle ground.

The findings thus strongly suggest a combination of the contact-based approach and the degree of institutionalization. According to Bao (2009, cf. section 2.4), in the formation of a new variety, the productivity of a transferred feature depends crucially on whether that feature violates constraints operative within the emergent variety. The emergent variety itself is heavily influenced by the superstrate, considering that only features which can be “expressed felicitously” by “the morphosyntactic materials” of the superstrate, do in fact surface in the emergent variety (ibid.: 347). Consequently, if a feature cannot be expressed by morphosyntactic material of the lexifier, the feature will not surface. If, however, it can be expressed felicitously, then its productivity depends on the degree to which it violates constraints in the new variety. As this study suggests, the constraints in the new varieties are influenced by the degree of institutionalization of these varieties. In the case of V>N conversion, the more institutionalized a variety is, the further the constraints operative in it seem to approximate those of native varieties. As a consequence, the further evolved a variety is, the closer its frequency pattern of V>N conversion is to that of native varieties such as BrE and USE. In other words, for optimal productivity, new features in non-native varieties have to be compatible with both the substratum and the superstratum. Verb-to-noun conversion, although structurally compatible with English, is not the preferred nominalization process (derivation is much more frequent), so that in highly institutionalized varieties such as SgE

this innovation is less frequent than in less institutionalized varieties such as HKE, where the influence of the substrate is greater.

The inclusion of IndE as a basis of comparison with largely synthetic substrata has yielded results that substantiate the claim that the degree of institutionalization is crucial in the development of new varieties. Contrary to what was expected, the frequency profile of V>N conversion in IndE does not resemble that of native varieties, even though substrate influence as an instigator for higher levels of V>N conversion can largely be ruled out. More to the point, looking at the picture for IndE, one could ask why the question of the influence of the substratum arises at all, if SgE and IndE show equally high odds of conversion and speakers react almost equally fast to it. Is the substrate language probably not at all relevant in shaping usage patterns of V>N conversion? A closer look at the regression coefficients as well as the qualitative analysis of select examples, however, reveals that the odds of conversion in SgE and IndE are arrived at through distinct, and different, pathways. In IndE, the blocking effect exerted by the frequency of the near-synonymous deverbal noun is smaller, constraining V>N conversion less, and thus resulting in higher odds of conversion. Furthermore, in IndE, word-formation processes in general can seemingly be applied with fewer restrictions, producing innovative formations more readily (as illustrated in section 7.3). In SgE, on the other hand, the odds of conversion appear to be due to transfer from the analytic substrate language. This can mainly be concluded from the similarities between SgE and HKE as regards the formal constraints, more precisely, from the fact that in both varieties V>N conversion preferably occurs in explicitly nominal contexts.

In summary, the results, both of the corpus-analytic and the experimental study, suggest that in the gradual expansion of an already existing lexical process, such as V>N conversion, the effect of the socio-institutional status of English—and of language proficiency, which results from it—is the more powerful explanatory factor than substrate influence. Particularly in the rating task, the acceptability of V>N conversion, along with other non-standard features, depended on whether participants used English every day or not. The sample of IndE participants largely consisted of speakers whose native language was not English and some of the IndE participants indicated that they did not use English on a daily basis. This resulted in a rating behavior which was in parts similar to that of the HKE participants. Since the two varieties do not share substrata, similarities are presumed to be rooted in similar sociolinguistic contexts, with English assuming comparable functions in both contexts.

That the usage pattern of V>N conversion in more institutionalized new varieties approximates that of native varieties is in line with Edwards and Laporte's (2015: 160) findings on the preposition *into*: "The more advanced a variety in Schneider's model (i.e. the more insti-

tutionalised), the more similar it was to ENL [English as a native language], while the least institutionalised varieties were the most distant from ENL.” That more advanced varieties show patterns more similar to the native varieties was also found by Deshors and Gries (2016), who analyzed *to*-infinitival and gerundial verb complements in the same varieties that this study scrutinizes. Nonetheless, these results are surprising in light of studies such as Mukherjee and Gries’s (2009), which found that in less evolved varieties verb-complementation patterns of in-, mono-, and ditransitive verbs are similar to native varieties, whereas more institutionalized varieties differ more from native varieties, presumably due to a higher endonormative orientation. This leads them to equate advanced development of an individual variety with increased distance from the parent variety. These seemingly contradictory results on verb-complementation patterns by Deshors and Gries (2016) on the one hand and Mukherjee and Gries (2009) on the other hand (even though both studies share the database, namely the ICE corpora) imply that “the evolution of World Englishes does not necessarily have the same impact on all linguistic features” (Laporte 2012: 286) and that different innovative features can be affected in different ways by ongoing institutionalization (cf. Bernaisch 2015: 214–218). Or, as Deshors and Gries (2016) formulate: “emancipation can, but need not always, result in unidirectional pathways away from the historical source variety”.

These results thus call for a refinement of the Dynamic Model, which accounts for the possibility that, first, features develop in different ways within the same variety, and, second, that individual features take one of two distinct developmental paths: either ‘away’ from the parent variety or ‘towards’ it. While the Dynamic Model in its original form does not exclude these possibilities, Schneider (2007: 45) explicitly states that “[w]here innovations originate [...] is not of primary importance in the long run”—to his model, one must add. This has resulted in e.g. Mukherjee and Gries (cf. 2009: 36) interpreting endonormative stabilization as the development of patterns and rules unique to the new variety and different from the parent variety. Nonetheless, it seems that endonormativity is better represented by conceptualizing it as consolidated competence and greater fluency in English, most likely paired with a greater self-reliance of the speakers of a variety. In such varieties, substrate transfer is less likely to be used by individuals as a compensatory learner strategy in ad-hoc and unpredictable ways. When substrate transfer is attested, it is likely to be in cases which represent community consensus. In other words, the general frequency of instances of substrate transfer may decline, but substrate transfer in locally conventionalized instances may stabilize nevertheless.

In the particular case of V>N conversion, this results in the more advanced variety exhibiting a usage pattern that is closest to the native varieties. Backup for this interpretation

comes from Bernaisch (cf. 2015: 218), whose study on Sri Lankan English reveals that exo- and endonormative tendencies can coexist within the same variety. In other words, the same variety can at the same time exhibit features that are more exonormative, i.e. similar to the native varieties, and features that are more endonormative, i.e. dissimilar from the native varieties. Or, more generally put, “[e]xo[-] and endonormative orientations, nativization, and differentiation can each work simultaneously in the development of a variety and do not have to be understood as separate, discrete stages” (Onysko 2016: 201). It is thus the “interaction” between these “forces” that crucially influences how varieties develop (ibid.).

This also explains why both V>N conversion as well as the preposition *into* do not follow the path predicted by Hoffmann (2014) for constructions in the Dynamic Model. According to Hoffmann (ibid.: 171–172), later developmental stages should see an increased productivity and higher type frequencies of innovative features due to an increasing abstraction. However, this is not the case for V>N conversion, which, according to the data presented here, is characterized by a higher type frequency in less advanced varieties (supposedly largely in instances of ad-hoc substrate transfer). It can therefore be concluded that there must be more than one possible trajectory which innovative features can follow in new varieties.³

In summary, as far as the Dynamic Model as a representation of the linguistic reality of Asian Englishes is concerned, the results of the present study strengthen Edwards and Laporte’s (2015: 162) claim that “Schneider’s model [...] may be more applicable when considering sociocultural aspects such as identity issues, but less so for investigating structural features in isolation”, considering that the Dynamic Model in its current state does not account for features developing at the same time both similarly and dissimilarly to the parent/native variety. Bernaisch (2015: 218) adds to this claim by asserting that “the process of structural nativisation as suggested by Schneider (2003, 2007) is a generalisation of the sub-processes occurring at the level of nativisation indicators”, where “[n]ativisation indicators are fine-grained units of language organisation on the basis of which endonormative and exonormative tendencies can be empirically investigated and modelled”.

Despite the criticism brought forward against the Dynamic Model in the preceding paragraphs, Schneider’s conceptualization of the emergence of new varieties of English should not be discarded carelessly. The Dynamic Model—with the aforementioned addition of one variety simultaneously containing exo- and endonormatively oriented features—still seems to be the most widely accepted model which most accurately explains V>N conversion in the Asian Englishes analyzed. Models which explain the formation and emergence of new

³For an extensive discussion of innovations in non-native varieties of English the reader is referred to a special issue of the *International Journal of Learner Corpus Research* (Vol. 2, No. 2, 2016).

varieties on purely structural grounds such as Bao's usage-based transfer from the substrate language fail to account for differences between varieties which are grounded in the different developmental stages of these varieties (here HKE vs. SgE). Another model unsuitable to explain V>N conversion in Asian varieties is Trudgill's (2004) approach to *new-dialect formation*, which explicitly denies the importance of identity in variety genesis (cf. Trudgill 2008).⁴

Newer models, which go beyond the notion of development and focus more on the globalization of English, such as Schneider's (2014b: 28) *Transnational Attraction* cannot explain the usage patterns of V>N conversion in Asian varieties either; V>N conversion is most likely not a case of "the appropriation of (components of) English(es) for whatever communicative purposes at hand, unbounded by distinctions of norms, nations or varieties", but rather an instance of transfer from the Chinese contact language. While the idea of *Transnational Attraction* might be appropriate in Expanding Circle contexts—for which it was originally conceived (cf. *ibid.*, for an example cf. Edwards 2016 on English in the Netherlands)—the conversion of verbs to nouns is hardly likely to be modeled on a native variety of English considering that the native varieties BrE and USE make use of V>N conversion to a much lower extent. As far as Mair's (2013a) *World System of Englishes*, which provides a hierarchical ordering of varieties in terms of global importance and thus influence, is concerned, it does not provide a better theoretical framework for V>N conversion either. While this cannot be excluded, it is extremely unlikely that the process should be appropriated from a less influential native variety of English in USE (cf. *ibid.* on the dominance of USE), or be transferred from USE to the Asian varieties in question.

Onysko (2016), however, argues that all World Englishes are instantiations of language contact (cf. *ibid.*: 205) and that the individual speaker is the "ultimate agen[t]" in language contact (*ibid.*: 211). Onysko's is a cognitive model of World Englishes, which comes at a time when cognitive linguistic approaches (such as Cognitive Construction Grammar or the usage-based paradigm in general) are on the rise. This model might thus be a first step towards reconciling the level of the individual speaker on the one hand, and the level of the community of speakers on the other. While the latter is of vital importance to and thus already well described in the Dynamic Model, the role of individual speakers' cognition in the emergence of World Englishes is yet to be explored and modeled. According to Onysko (*ibid.*: 209–212), there are three dimensions of variation in the emergence of contact Englishes: the "setting of the contact situation", the "processes of language contact", and the "parameters of language contact". The extra-linguistic conditions (e.g. "history and duration of contact",

⁴Cf. Schneider (2008: 266) on why Trudgill's view cannot be reconciled with the Dynamic Model.

“codes and cultures in contact”) as well as the mode and medium of contact and its “textual and contextual embedding” are subsumed under the setting. By processes, Onysko refers to cognitive and systemic processes which give rise to different contact phenomena (e.g. “analogical selection of linguistic units” leads to “transfer”). The “parameters comprise linguistic and extra-linguistic factors that play a decisive role for manifestations of language contact” such as the “typological overlap of the languages”, attitudes towards them or language policies (Onysko 2016: 211). Yet, the most important of all parameters is “a speaker’s individual attitude towards language behavior”, as speakers are the “ultimate agen[ts]” in language contact situations (*ibid.*).

Five different types of prototypical contact Englishes emerge from these three dimensions of variation, depending on the specific combination of setting, processes and parameters. These types are global Englishes, learner Englishes, Englishes in multilingual constellations, English-based pidgins and creoles, and koiné Englishes (*cf. ibid.*: 212–214).⁵ An example for a variety of English in a multilingual constellation would be SgE, which is used on a par with Mandarin and other official mother tongues in daily life.

However, in this model, Englishes are not conceptualized as stative, but as dynamic varieties that can shift from one type to another, depending on the “intensity of contact at [the] time of formation” and “at [the] present time of use” (*ibid.*: 213–214). This means that Onysko’s model allows for a dynamic, circular, non-hierarchical classification of Englishes, where varieties can be re-allocated under the descriptor of a different variety type (*cf. ibid.*: 215). This supersedes Schneider’s linear conceptualization of the development of varieties of English and thus accommodates developments which the Dynamic Model cannot describe with sufficient accuracy, e.g. in cases where varieties ‘jump’ certain stages (e.g. no colonial past as in Namibia (*cf. Buschfeld 2014: 189*) or HKE, for which Görlach (*cf. 2002: 109*) predicts a development from an ESL variety back to an EFL variety due to the increasing influence of Mainland China and Mandarin Chinese). This model thus connects all types of varieties and emphasizes the gradience of variety types as well as the similarities between the cognitive processes underlying the variety prototypes. The conceptualization of prototypical categories further allows for varieties to be “peripheral members” of categories or to exhibit features characteristic of two prototypes (*cf. Onysko 2016: 215*).

Assuming that varieties can belong to more than one category and that they can switch categories helps explain the results of this study. While HKE, IndE, and SgE can certainly all be classified as Englishes in multilingual constellations, HKE (and to some extent also IndE) is not a prototypical member of this category but also shows characteristics typical

⁵The idea of a variety prototype is already mentioned in (Biewer 2011: 28), yet only for ESL varieties.

of a learner English. Consequently, according to this model, the categories of ESL and EFL varieties do not exclude each other—as has often been implied when referring to the tripartite categorization of varieties into EFL-ESL-ENL varieties—but rather share underlying cognitive processes and parameters (cf. *ibid.*). For example, in the case of Englishes in multilingual constellations (traditionally often called ESL varieties) and learner Englishes (or EFL varieties), the “[a]nalogical selection of linguistic units” in the L1 or any other frequently used language (e.g. Chinese in Hong Kong) can always result in transfer from that language, regardless of the context in which English is acquired and used, which in turn could result in the same linguistic feature surfacing in both variety types.

While Onysko’s model is certainly a right step in the direction of a cognitive conceptualization of language contact in which speakers are at the core and in which variety types are gradient and varieties in themselves dynamic entities, the cognitive processes mentioned in the model still have to be defined more rigorously and described for individual features. A case in point are simplification phenomena, which are well attested for English-based pidgins and creoles, learner Englishes, and also Englishes in multilingual constellations, but which do not feature in Onysko’s (*ibid.*: 210) description of “observable contact phenomena”.

9.4 Asian ESL varieties

What has already been touched upon in the preceding paragraphs but deserves further attention is the fact that the classification of varieties according to their historical context into ENL-ESL-EFL varieties is highly problematic. While all Asian varieties under scrutiny in this study were termed ESL varieties in the past, they show markedly different profiles for V>N conversion. Numerous studies (Biewer 2011; Deshors 2014; Edwards and Laporte 2015; Gilquin 2015; Gilquin and Granger 2011; Laporte 2012; Williams 1987; also cf. Gilquin 2015 for a commented list of prior studies) have previously insisted that the boundary between EFL and ESL varieties is “blurry” and that the distinction between EFL and ESL should rather be understood as a continuum. However, the inadequacy of the notion of *ESL variety* in itself has not been stressed enough. In focusing on three ‘classic’ ESL varieties, the present study has revealed that the notion of ESL variety should in itself be conceptualized as a continuum, accommodating different types of ESL varieties, some located closer to the ENL pole of the continuum (e.g. SgE) and others showing traces characteristic of the EFL pole (e.g. HKE). The picture that emerges for ESL varieties thus resembles second language acquisition (SLA) in certain respects. This is particularly the case for varieties which are less well indigenized, i.e. those that are located near the EFL pole, such as HKE, and whose speakers are thus not as

proficient as speakers of varieties closer to the ENL pole. In all ESL varieties, patterns from the L1 tend to be transferred to English, but an increasing proficiency in English inhibits and reduces this transfer process. More proficient speakers, e.g. of IndE or SgE, are able to deal with more complex and more schematic constructions, which is why the likelihood of occurrence of derivation in SgE and IndE is higher than in HKE, while in HKE, conversion is comparatively more likely than in SgE and IndE. This again points towards the greater importance of the degree of institutionalization in shaping contact varieties compared to the often over-estimated influence of substrata.

The similarities between ESL and SLA also reflect in language processing (cf. section 8.9.2). As Laporte (2012: 287) suggests, “similar developmental and cognitive processes are perhaps at play across both EFL and ESL acquisition”. Deshors (2014: 300) and Deshors and Gries (2014: 201) found that in EFL contexts, speakers resort to processes familiar to them from their L1 in “complex grammatical contexts”, that is, in situations with a “higher cognitive load” (Deshors 2014: 300). The results of the maze task indicate that the same phenomenon can be observed in HKE. HKE speakers, more than any other group of participants, reacted faster to the stimuli with the converted forms, which result from a process that is very unconstrained in the HKE speakers’ L1. In this situation, which can be assumed to impose a high cognitive burden on the participants, HKE speakers were fastest to react to V>N conversion, supposedly because of their falling back on their L1. The analysis of the spoken corpus data presented in section 7.1.2 points in the same direction.

This study has thus contributed to bridging the paradigm gap which has often been invoked to describe the fact that even though learner varieties and second-language varieties share many features and processes they have (or had) not been studied from a comparative perspective (cf. Mukherjee and Hundt 2011; Sridhar and Sridhar 1986). This study adds to the line of research showing that ENL, ESL and EFL varieties do indeed share characteristics and that it makes sense to envisage second-language varieties as a special type of learner languages (or, in Onysko’s terms, as peripheral members of the learner Englishes category). The New Englishes are ultimately also learner Englishes, although “these varieties [...] can no longer be considered learner varieties” but “have become regional standards” (Williams 1987: 163).

On a different note, it has to be pointed out that—not unlike the notion of *ESL variety*—notions such as *Asian Englishes* or *Chinese Englishes* as terms to designate groups of varieties must be rejected because they simply are too broad to accurately represent the linguistic reality. These terms generally do not reflect specific cultural and sociolinguistic circumstances well enough. As Leimgruber (2013c: 6) notes: “Such [geographic or political] labels suggest

a certain degree of uniformity within the variety which is often lacking.” The problematic nature of the “attachment of vague labels to ill-defined regional varieties of English” has also been stressed for other groups of varieties such as the ‘Celtic Englishes’ (Görlach 1997: 27). As Görlach (*ibid.*: 46) asserts, labels such as these are often merely “scholarly construct[s]”.

9.5 Processing V>N conversion

Conversion is a process that speakers of varieties with a Chinese substratum know from their L1, as has previously been pointed out. Furthermore, it is also a process that is often associated with learner varieties (cf. Pavese 1998: 215). Conversion results in a regularization of the paradigm and decreases redundancy—at the expense of potentially over-increasing ambiguity. While conversion may thus present an advantage in encoding in that the speaker does not have to retrieve the derived form, it may also become a disadvantage in decoding. If the converted form is embedded in an ambiguous context, it can take the hearer even longer to decode the meaning, that is, to coerce the DEVERBAL CONVERTED NOUN construction. Conversion was therefore assumed to occur predominantly in explicitly nominal contexts. In these contexts, conversion is predicted to result in a minimized processing effort on the part of the hearer, as coercion is facilitated by the explicitness of the context. The qualitative analysis of corpus data from the Asian varieties presented in chapter 7 has revealed that conversion often occurs embedded in explicitly nominal contexts in HKE and SgE, but not to the same extent in IndE.

As has been shown above, it appears that speakers of more advanced varieties profit less from the reduced processing effort of conversion and instead prefer the derived form. This can be expected to be rooted not only in the fact that for more advanced or native speakers derived forms are easily accessible, but also in hearer-orientation on the part of the speaker. Even though conversion can be considered the linguistically more economic process for the speaker compared to derivation—in line with Hawkins’s (2004) principles—it comes at the expense of the hearer having to coerce the construction. The aims of producing language as effectively as possible while at the same time making oneself understood counteract each other. If speakers are able to draw on other constructions that do not require the hearer to coerce the construction in order to understand it, they might use this construction rather than the DEVERBAL CONVERTED NOUN construction. For the (more) nativized varieties, derivation is the construction that is more easily processable and thus preferable to conversion (cf. chapters 6 and 8).

For HKE, the excerpts in section 7.1.2 indicate that speakers make use of conversion particularly in situations with a presumably high processing burden. The excerpts further show that the interlocutors do not react negatively to the converted forms, neither initiating repair nor asking for clarification. On the contrary, they provide backchanneling tokens, which encourage the speaker to continue (e.g. *right* in 7.7, *aw* meaning ‘oh I see’ in 7.8). Consequently, the cognitive burden on the hearers in perceiving and decoding V>N conversion seems comparatively low. This might hold for speakers of all varieties or could be particular to the HKE hearers, who, owing to their higher familiarity with this process, coerce the DEVERBAL CONVERTED NOUN construction faster. Yet, this claim can only be made with reservations; a larger database of spoken data would be necessary to substantiate it.

9.6 A note on methodology

Approaching verb-to-noun conversion from a Construction Grammar perspective has turned out to be highly beneficial to an account of conversion. The view that has been adopted does not restrict itself to either a purely morphosyntactic or a purely lexical take on the phenomenon, but understands verb-to-noun conversion as a construction in its own right. This has helped explain the findings for the individual varieties. In varieties which are similar to learner varieties, mostly HKE, conversion is favored to a greater extent than in more nativized varieties. From a Construction Grammar perspective, the explanation for this is readily at hand: V>N conversion looks like a substantive and atomic construction (even though it is not), and when it occurs embedded in explicitly nominal contexts, the processing cost required to coerce the construction is minimal. In contrast, in more advanced or in native varieties, more complex and schematic constructions are more productive and more common, as Hoffmann (2014: 171–172, 174) argues. Derivation is such a complex, partly schematic construction. While it may be more difficult to process for speakers of HKE, it is less of a challenge for speakers of SgE and IndE, the more nativized varieties, which explains why speakers of these varieties prefer it over conversion to a greater extent compared to speakers of HKE.

The method of combining corpus-linguistic and experimental data which has been adopted in this study has proven fruitful in shedding light on complex linguistic settings such as the ones encountered in the Asian context. The corpus analysis has brought to light distinct usage patterns for V>N conversion, which have been consolidated by data on the acceptability of V>N conversion. This has resulted in a coherent picture in that in varieties where conversion is more frequent, the phenomenon is also judged more acceptable. In addition, the

experimental study has provided insights into the processing of conversion, which also tally with the previous findings: Those participants who speak varieties in which V>N conversion has been found to occur more often also react to V>N conversion more quickly compared to speakers of varieties in which verb-to-noun conversion is infrequent.

Corpus analytic and experimental methods thus complement each other, yielding a deeper insight into the usage-based nature of language in general, and verb-to-noun conversion in particular. This combination should therefore be applied to investigate a range of other linguistic phenomena and processes, as has been and is currently done for e.g. the grammaticalization of modal verbs (cf. Lorenz 2013) and selected simplification processes (cf. Terassa in preparation).

As far as the corpus-linguistic methodology is concerned, a combination of quantitative and qualitative analyses has proven extremely rewarding. Tendencies visible in the regression models, such as the comparatively high odds of conversion in IndE or the comparatively weaker blocking constraint, could not have been interpreted in a meaningful manner had a qualitative analysis not been undertaken. A quantitative analysis complemented by an in-depth qualitative analysis could also be applied to similar language phenomena, for example to investigate in how far the effects found for conversion and derivation apply to other conditions where analytic, non-morphemic forms and near-synonymous, synthetic forms interact or even compete. A case which comes to mind is that of demonyms, i.e. of comparing the use of analytic and synthetic formations such as *Hong Kong people* vs. *Hong Kongers* (cf. Chen 2016). Studies such as Chen's (ibid.) require expanding the scope and taking into account cultural influences on language.

Finally, a critical note on a statistical account of varieties of English is in order. As has been suggested above, the Dynamic Model does not straightforwardly predict the usage pattern of V>N conversion in new varieties of English. Yet, disproving the Dynamic Model (or any other of the models presented above) on statistical grounds is impossible considering that cultural aspects such as (linguistic) "identity constructions" or the "sociopolitical background"—both key to the development of varieties according to the Dynamic Model (cf. Schneider 2007: 29–55)—defy quantification. Moreover, as this study has shown, frequency is just one of various factors in a complex multifactorial network. For such research questions, quantitative approaches may serve as a useful control on qualitative interpretation or help produce a more fine-grained picture of selected aspects. An exclusively quantitative account would be reductionist. Here, a quantitative take on V>N conversion in varieties in the form of multivariate statistics has contributed to refining the Dynamic Model in its current form.

9.7 Conclusion

The study in hand set out to explore verb-to-noun conversion in Asian varieties of English. In order to pursue this goal, a Cognitive Construction Grammar approach, situated within the usage-based paradigm, was adopted. V>N conversion is understood as the embedding of a verb in a nominal frame. To make sense of this construction, the hearer needs to coerce it. Therefore, conversion was hypothesized to occur mostly in explicitly nominal contexts such as the NOUN PHRASE construction, which reduce the processing effort of V>N conversion. This was found to be the case in almost all of the varieties investigated, with IndE constituting a notable exception (cf. chapter 7). By viewing verb-to-noun conversion as a construction, it was possible to bridge the gap between lexis and grammar, thus investigating conversion without the need to focus on whether it should be ascribed to the domain of lexis or morphosyntax. The emergence of V>N conversions was scrutinized in detail in USE, the most influential of all varieties (cf. Mair 2013b: 261). The analysis of the evolution of DISCONNECT shed light on the formal and functional as well as semantic and stylistic developments that come with a change in word class (cf. chapter 5).

Integrating a Construction Grammar approach with the evolution of varieties of English along the lines of Hoffmann's (2014) work proved to be challenging. Even though previous research (cf. e.g. Mukherjee and Gries 2009) has illustrated that more advanced varieties "exhibit a greater type frequency" of innovative constructions due to a deeper entrenchment of these (Hoffmann 2014: 172), this is not the developmental path observed for verb-to-noun conversion. On the contrary, more advanced varieties show less verb-to-noun conversion, thus approximating native varieties—at least on the surface. At a deeper level and for other features, locally conventionalized deviations from the old colonial norm may exist and therefore still signal endonormativity (e.g. verb-complementational profiles). It is therefore necessary to assume that different constructions can take different developmental paths, some displaying endonormative tendencies, some exonormative tendencies (cf. Bernaisch 2015: 218). In the course of their development, some innovations may solidify into robust features of a new variety, following new local norms, while others become more infrequent and may eventually be lost. The particular path a construction takes will depend on the complex interaction of the influence of substrate languages and the degree of institutionalization of English.

In pursuing a usage-based approach to verb-to-noun conversion the present study has contributed to exploring the explanatory power of frequency as a factor in language contact and language processing. Three effects deserve particular attention. First, the frequency of a

verbal base predicts the odds of a verb being converted to a noun. In USE, a high frequency of the verbal base relative to the converted form was discovered to block conversion, as in the case of *CONNECT* (cf. chapter 5). In HKE, a decreasing absolute frequency of the verbal base increases the odds of conversion to a noun, while more frequent verbs prompt lower odds of conversion (cf. chapter 6). Second, the frequency of a near-synonym crucially influences the odds of conversion of a verb to a noun, with higher frequencies resulting in a stronger blocking effect, that is, in decreasing odds of conversion. The blocking effect was found to hold in all varieties investigated, albeit to different degrees. Third, a relation was observed to exist between how often verb-to-noun conversion occurs in the GloWbE corpus and how acceptable speakers judge it to be and how fast they react to it (cf. chapter 8). Speakers' experience with verb-to-noun conversion, operationalized by frequency of occurrence in corpora, proved to be a significant predictor of speakers' acceptability judgments, with speakers of varieties in which V>N conversion is encountered more frequently judging V>N conversion as more acceptable. As regards reaction times, it was shown that speakers with a higher familiarity with the process (as indicated by a higher frequency of occurrence in corpora) are significantly faster at processing V>N converted forms.

As far as the methodology is concerned, the availability of the large GloWbE corpus was essential for this project. V>N conversion is rather infrequent and without vast amounts of data investigating the phenomenon from a quantitative perspective would have been impossible. Nonetheless, combining the large GloWbE corpus with the smaller and more neatly compiled ICE corpora was rewarding, as the latter facilitated an investigation into register differences between varieties, which is currently not possible with GloWbE. As regards GloWbE as a new resource in World Englishes studies, the results demonstrate that large amounts of data cancel out noise unavoidable in comparatively untidy sampling procedures, i.e. where data are obtained (semi-)automatically with limited manual post-editing.

Not only the combination of small and large corpora but also the integration of corpus analytic and experimental methods proved to be valuable for both research within the usage-based paradigm as well as research into World Englishes. The results of the web-based experiment confirmed the results of the corpus analysis, showing that this combination of methods is a suitable way of empirically investigating Schmid's (2000: 39) *From-Corpus-to-Cognition Principle*.

The success of web-based experimentation in the present study encourages a further and more systematic application of this method. Particularly in the field of World Englishes, where researchers are often at a long distance from speakers of the varieties in question, web-based experimentation is a time-efficient and resource-friendly way of obtaining native

speaker data. As this project suggests, in experimentation (as in corpora), large amounts of data “swamp” (McGraw et al. 2000: 505) noise introduced by less careful participant selection and monitoring during the experiment.

The study in hand has thus opened doors for further research on word-formation processes in World Englishes. For example, it would be worth exploring whether the tendencies observed here also hold for other varieties as well as for other word-formation processes or other directions of conversion. Generally, many additional topics of interest hinge on the availability of suitable corpora. The publication of a revised version of GloWbE with a more fine-grained classification of the web registers in the near future (cf. Davies 2015) will provide more insights into register-specific aspects of conversion. Diachronic corpora of varieties of English could help pursue further investigation into the developmental stages of English, e.g. by answering the question of whether verb-to-noun conversion in SgE at an earlier stage resembled verb-to-noun conversion in HKE in its current state.⁶ A systematic comparison of the spoken and the written mode or of the basilectal and acrolectal variants of varieties (e.g. Singlish vs. Standard Singapore English) could also lead to a more detailed understanding of the trajectory of innovations as well as of the role of the substrate in the emergence of varieties more generally. Also, an analysis of a specific group of semantically related verbs could be of interest, in order to unearth the precise role of semantics in the success of conversion. Moreover, a more detailed definition of the *DEVERBAL CONVERTED NOUN* construction is in order, particularly as regards its pragmatic meaning, which should also be contrasted with the meanings of near-synonymous constructions such as derivation or the *LIGHT VERB* construction.

In summary, this study constitutes a point of departure for future usage-based research into variation at the lexis-grammar interface in varieties of English. The availability of mega-corpora such as GloWbE as well as of web-based experimentation facilitates statistical modeling of differences between varieties. If the path taken in this study is pursued by future research and thus extended to cover more features and more varieties, a much deeper understanding of the evolution of constructions in different varieties of English lies ahead.

⁶Such corpora are currently being compiled by Biewer and colleagues for HKE (Biewer 2016; Biewer et al. 2014) and Hoffmann and colleagues for SgE (Hoffmann et al. 2012).

Bibliography

- Abbey, Susan, David Christie, Barbara Derkow-Disselbeck, Laurence Harger, and Allen J. Woppert. 2011. *English G 21, A4*. Berlin: Cornelsen.
- Akaike, Hirotugu. 1973. 'Information theory and an extension of the maximum likelihood principle.' *Proceedings of the 2nd International Symposium on Information Theory*. Ed. F. Csáki and Petrov B. N. Budapest: Akadémiai Kiado. 267–281.
- Akaike, Hirotugu. 1974. 'A new look at the statistical model identification.' *IEEE Transactions on Automatic Control* AC-19.6: 716–723.
- Alsagoff, Lubna. 2010. 'English in Singapore: Culture, capital and identity in linguistic variation.' *World Englishes* 29.3: 336–348.
- Alsagoff, Lubna. 2012. 'The development of English in Singapore. Language policy and planning in nation building.' *English in Southeast Asia. Features, policy and language in use*. Ed. Ee-Ling Low and Azirah Hashim. Amsterdam: John Benjamins. 137–154.
- American National Corpus Project. 2012a. *About the ANC project*. <<http://www.anc.org/about/>> (accessed December 3, 2014).
- American National Corpus Project. 2012b. *ANC second release*. <<http://www.anc.org/data/anc-second-release/>> (accessed December 3, 2014).
- Ansaldo, Umberto. 2004. 'The evolution of Singapore English. Finding the matrix.' *Singapore English. A grammatical description*. Ed. Lisa Lim. Amsterdam: John Benjamins. 127–149.
- Arnon, Inbal and Neal Snider. 2010. 'More than words: Frequency effects for multi-word phrases.' *Journal of Memory and Language* 62.1: 67–82.
- Aronoff, Mark. 1976. *Word formation in Generative Grammar*. Cambridge: MIT Press.
- Ashford, Stephanie, Paul Aston, and Rosemary Hellyer-Jones. 2000. *Green Line New 4*. Stuttgart: Klett.
- Baayen, R. Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baayen, R. Harald. 2009. 'Corpus linguistics in morphology: Morphological productivity.' *Corpus linguistics. An international handbook*. Ed. Anke Lüdeling and Merja Kytö. Vol. 2. Berlin: De Gruyter. 900–919.

Bibliography

- Baayen, R. Harald and Petar Milin. 2010. 'Analyzing reaction times.' *International Journal of Psychological Research* 3.2: 12–28.
- Baeskow, Heike. 2006. 'Reflections on noun-to-verb conversion in English.' *Zeitschrift für Sprachwissenschaft* 25: 205–237.
- Balteiro, María Isabel. 2001. 'On the status of conversion in present-day American English. Controversial issues and corpus-based study.' *Atlantis* 23.2: 7–29.
- Balteiro, María Isabel. 2007a. *A contribution to the study of conversion in English*. Münster: Waxmann.
- Balteiro, María Isabel. 2007b. *The directionality of conversion in English. A dia-synchronic study*. Bern: Peter Lang.
- Bao, Zhiming. 2005. 'The aspectual system of Singapore English and the systemic substratist explanation.' *Journal of Linguistics* 41.2: 237–267.
- Bao, Zhiming. 2009. 'One in Singapore English.' *Studies in Language* 33.2: 338–365.
- Bao, Zhiming. 2010a. 'A usage-based approach to substratum transfer. The case of four unproductive features in Singapore English.' *Language* 86.4: 792–820.
- Bao, Zhiming. 2010b. 'Must in Singapore English.' *Lingua* 120.7: 1727–1737.
- Bao, Zhiming. 2011. 'Convergence-to-substratum and the passives in Singapore English.' *Creoles, their substrates, and language typology*. Ed. Claire Lefebvre. Amsterdam: John Benjamins. 253–270.
- Baroni, Marco, Emiliano Guevara, and Roberto Zamparelli. 2009. 'The dual nature of deverbal nominal constructions: Evidence from acceptability ratings and corpus analysis.' *Corpus Linguistics and Linguistic Theory* 5.1: 27–60.
- Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2014. *lme4: Linear mixed-effects models using Eigen and S4. R package Version 1.1-7*. <<http://cran.r-project.org/package=lme4>>.
- Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.
- Bauer, Laurie. 2002. *English word-formation*. Repr. Cambridge: Cambridge University Press.
- Bauer, Laurie. 2003. *Introducing linguistic morphology*. 2nd ed. Washington: Georgetown University Press.
- Baumgardner, Robert J. 1998. 'Word-formation in Pakistani English.' *English World-Wide* 19.2: 205–246.
- Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. 'Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys.' *American Journal of Political Science* 58.3: 739–753.

- Bernaisch, Tobias. 2015. *The lexis and lexicogrammar of Sri Lankan English*. Amsterdam: John Benjamins.
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, Douglas. 1989. 'A typology of English texts.' *Linguistics* 27.1: 3–43.
- Biber, Douglas. 1993. 'Representativeness in corpus design.' *Literary and Linguistic Computing* 8.4: 243–257.
- Biber, Douglas. 2003. 'Compressed noun-phrase structures in newspaper discourse: the competing demands of popularization vs. economy.' *New media language*. Ed. Jean Aitchison and Diana M. Lewis. London: Routledge. 169–181.
- Biber, Douglas. 2010. 'Corpus-based and corpus-driven analyses of language variation and use.' *The Oxford handbook of linguistic analysis*. Ed. Bernd Heine and Heiko Narrog. Oxford: Oxford University Press. 159–192.
- Biber, Douglas, Jesse Egbert, and Mark Davies. 2015. 'Exploring the composition of the searchable web: A corpus-based taxonomy of web registers.' *Corpora* 10.1: 11–45.
- Biber, Douglas and Jerry Kurjian. 2007. 'Towards a taxonomy of web registers and text types: A multi-dimensional analysis.' *Corpus linguistics and the web*. Ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. Amsterdam: Rodopi. 109–131.
- Bickerton, Derek. 1981. *Roots of language*. Ann Arbor: Karoma.
- Biermeier, Thomas. 2008. *Word-formation in New Englishes. A corpus-based analysis*. Münster: Lit.
- Biewer, Carolin. 2011. 'Modal auxiliaries in second language varieties of English: A learner's perspective.' *Exploring second-language varieties of English and learner Englishes. Bridging a paradigm gap*. Ed. Joybrato Mukherjee and Marianne Hundt. Amsterdam: John Benjamins. 7–34.
- Biewer, Carolin. 2016. *Anglistik und Amerikanistik: Biewer*. <http://www.anglistik.uni-wuerzburg.de/abteilungen/englische_sprachwissenschaft/mitarbeiter_innen/biewer/> (accessed August 24, 2016).
- Biewer, Carolin, Tobias Bernaisch, Mike Berger, and Benedikt Heller. 2014. 'Compiling *The Diachronic Corpus of Hong Kong English* (DC-HKE): Motivation, progress and challenges.' ICAME 35. Nottingham, April 30, 2014. <http://www.ling.arts.kuleuven.be/qlvl/prints/Biewer_Bernaisch_Berger_Heller_2014pres_Compiling_DCHKE.pdf>.
- Birnbaum, Michael H. 2004a. 'Human research and data collection via the internet.' *Annual Review of Psychology* 55: 803–832.

Bibliography

- Birnbaum, Michael H. 2004b. 'Methodological and ethical issues in conducting social psychology research via the internet.' *The Sage handbook of methods in social psychology*. Ed. Carol Sansone, Carolyn C. Morf, and A. T. Panter. Thousand Oaks: Sage Publications. 359–382.
- Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories. What corpus data do and do not reveal about the mind*. Berlin: De Gruyter.
- Boas, Hans C. 2011. 'Coercion and leaking argument structures in Construction Grammar.' *Linguistics* 49.6: 1271–1303.
- Bock, J. Kathryn. 1986. 'Syntactic persistence in language production.' *Cognitive Psychology* 18.3: 355–387.
- Bolton, Kingsley. 2000. 'The sociolinguistics of Hong Kong and the space for Hong Kong English.' *World Englishes* 19.3: 265–285.
- Bolton, Kingsley, ed. 2002. *Hong Kong English. Autonomy and creativity*. Hong Kong: Hong Kong University Press.
- Bolton, Kingsley. 2003. *Chinese Englishes. A sociolinguistic history*. Cambridge: Cambridge University Press.
- Bolton, Kingsley. 2012. 'Language policy and planning in Hong Kong. The historical context and current realities.' *English in Southeast Asia. Features, policy and language in use*. Ed. Ee-Ling Low and Azirah Hashim. Amsterdam: John Benjamins. 221–238.
- Bolton, Kingsley and Bee Chin Ng. 2014. 'The dynamics of multilingualism in contemporary Singapore.' *World Englishes* 33.3: 307–318.
- Bortz, Jürgen. 2005. *Statistik für Human- und Sozialwissenschaftler*. 6th ed. Heidelberg: Springer.
- Boyd, Jeremy K. and Adele E. Goldberg. 2011. 'Learning what NOT to say. The role of statistical preemption and categorization in *a*-adjective production.' *Language* 87.1: 55–83.
- Brandt, Silke and Evan Kidd. 2011. 'Relative clause acquisition and representation. Evidence from spontaneous speech, sentence repetition, and comprehension.' *Converging evidence. Methodological and theoretical issues for linguistic research*. Ed. Doris Schönefeld. Vol. 33. Human cognitive processing. Amsterdam: John Benjamins. 273–291.
- Bresnan, Joan. 2007. 'A few lessons from typology.' *Linguistic Typology* 11.1: 297–306.
- Brown, Penelope and Stephen C. Levinson. 1987. *Politeness. Some universals in language usage*. Cambridge: Cambridge University Press.
- Bunton, David. 1991. 'A comparison of English errors made by Hong Kong students and those made by non-native learners of English internationally.' *Institute of Language in Education Journal Special Issue 2*: 9–22.

- Buschfeld, Sarah. 2013. *English in Cyprus or Cyprus English. An empirical investigation of variety status*. Amsterdam: John Benjamins.
- Buschfeld, Sarah. 2014. 'English in Cyprus and Namibia. A critical approach to taxonomies and models of World Englishes and second language acquisition research.' *The evolution of Englishes. The Dynamic Model and beyond*. Ed. Sarah Buschfeld, Thomas Hoffmann, Magnus Huber, and Alexander Kautzsch. Amsterdam: John Benjamins. 181–202.
- Buschfeld, Sarah, Thomas Hoffmann, Magnus Huber, and Alexander Kautzsch, eds. 2014. *The evolution of Englishes. The Dynamic Model and beyond*. Amsterdam: John Benjamins.
- Bybee, Joan L. 2006. 'From usage to grammar: The mind's response to repetition.' *Language* 82.4: 711–733.
- Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Callies, Marcus. 2015. 'Towards a process-oriented approach to comparing EFL and ESL varieties: A corpus-study of lexical innovations.' pre-conference workshop 'Corpus linguistics and linguistic innovations in non-native Englishes' at ICAME 36. Trier, May 27, 2015.
- Cannon, Garland. 1985. 'Functional shift in English.' *Linguistics* 23: 411–431.
- Census and Statistics Department, Hong Kong Special Administrative Region. 2013. *Thematic Household Survey - Report No. 51. Use of language in Hong Kong. Utilisation of child health and family planning services provided by maternal and child health centres*. Hong Kong. <<http://www.statistics.gov.hk/pub/B11302512013XXXXB0100.pdf>> (accessed March 26, 2014).
- Census of India. n.d.(a). *General note*. Ed. Government of India, Ministry of Home Affairs, Office of the Registrar General & Census Commissioner, India. <http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/gen_note.html> (accessed July 29, 2014).
- Census of India. n.d.(b). *Statement 6. Comparative rankings of scheduled languages in descending order of speaker's strength - 1971, 1981, 1991 and 2001*. Ed. Government of India, Ministry of Home Affairs, Office of the Registrar General & Census Commissioner, India. <http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement6.aspx> (accessed October 28, 2014).
- Census of India. n.d.(c). *Statement 8. Growth of non-scheduled languages 1971, 1981, 1991 and 2001 [sic]*. Ed. Government of India, Ministry of Home Affairs, Office of the Registrar General & Census Commissioner, India. <http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement8.aspx> (accessed July 29, 2014).
- Census of India. n.d.(d). *Statement 9. Family-wise grouping of the 122 scheduled and non-scheduled languages -2001*. Ed. Government of India, Ministry of Home Affairs, Office of the

Bibliography

- Registrar General & Census Commissioner, India. <http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/statement9.aspx> (accessed July 29, 2014).
- Chan, Alice H. D., Kang-Kwong Luke, Ping Li, Virginia Yip, Geng Li, Brendan Weekes, and Li Hai Tan. 2008. 'Neural correlates of nouns and verbs in early bilinguals.' *Annals of the New York Academy of Sciences* 1145: 30–40.
- Chan, Alice Y. W. 2010. 'Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners.' *TESOL Quarterly* 44.2: 295–319.
- Chan, Jim Y. H. 2013. 'Contextual variation and Hong Kong English.' *World Englishes* 32.1: 54–74.
- Chen, Wei. 2016. *Identity and diversity in Hong Kong: Changing linguistic naming conventions*. M.A. thesis. Freiburg: Albert-Ludwigs-Universität.
- Cheshire, Jenny. 2004. 'Sex and gender in variationist research.' *The handbook of language variation and change*. Ed. J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes. Oxford: Blackwell. 423–443.
- China Labor News Translations. n.d. *About*. <<http://www.cntranslations.org/about/>> (accessed February 12, 2015).
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.
- Chui, Sze Yan. 2010. 'Sensitivity to differences between speech and writing: Hong Kong students' use of syntactic features in English.' PhD thesis. Hong Kong: The Chinese University of Hong Kong.
- Clark, Eve V. and Herbert H. Clark. 1979. 'When nouns surface as verbs.' *Language* 55.4: 767–811.
- Collins, Peter and Xinyue Yao. 2013. 'Colloquial features in World Englishes.' *International Journal of Corpus Linguistics* 18.4: 479–505.
- Columbus, Georgie. 2010. *ICE-Canada. Codes for metadata*. <http://ice-corpora.net/ice/downloads/Codes_for_ICE_CAN_metadata.pdf> (accessed October 26, 2015).
- Cook, Paul and Graeme Hirst. 2012. 'Do web corpora from top-level domains represent national varieties of English?' *11es Journées internationales d'analyse statistique des données textuelles*. Ed. Anne Dister, Dominique Longrée, and Gérard Purnelle. 281–293. <<http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Cook,%20Paul%20et%20al.%20-%20Do%20web%20Corpora%20from%20Top-Level%20Domains.pdf>> (accessed January 21, 2015).
- Crawfurd, John. 1852. *A grammar and dictionary of the Malay language with a preliminary dissertation. Vol. I. Dissertation and grammar*. London: Smith, Elder, and Co. <<http://>

- wallace-online.org/converted/pdf/1852_Crawfurd_WS5.1.pdf> (accessed September 3, 2015).
- Crawley, Michael J. 2013. *The R book*. 2nd ed. Chichester: Wiley.
- Crump, Matthew J. C., John V. McDonnell, and Todd M. Gureckis. 2013. 'Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research.' *PLoS ONE* 8.8: 1–18.
- Crystal, David. 2004. 'English word classes.' *Fuzzy grammar. A reader*. Ed. Bas Aarts, David Denison, Evelien Keizer, and Gergana Popova. Oxford: Oxford University Press. 191–211.
- Crystal, David. 2008. 'Two thousand million?' *English Today* 24.1: 3–6.
- Crystal, David. 2011. *Internet linguistics. A student guide*. London: Routledge.
- Cysouw, Michael. 2014. *Languoid, doculect and glossonym: Formalizing the notion 'language'*. <<http://dlc.hypotheses.org/623>> (accessed January 26, 2015).
- Cysouw, Michael and Jeff Good. 2013. 'Languoid, doculect and glossonym: Formalizing the notion 'language'.' *Language Documentation and Conservation* 7: 331–359.
- Dako, Kari. 2001. 'Ghanaianisms. Towards a semantic and formal classification.' *English World-Wide* 22.1: 23–53.
- Dandurand, Frédéric, Thomas R. Shultz, and Kristine H. Onishi. 2008. 'Comparing online and lab methods in a problem-solving experiment.' *Behavior Research Methods* 40.2: 428–434.
- Das, Gurcharan. 2002. *The elephant paradigm. India wrestles with change*. New Delhi: Penguin Books.
- Davies, Mark. 2008–. *The Corpus of Contemporary American English: 450 million words, 1990–present*. <<http://corpus.byu.edu/coca/>>.
- Davies, Mark. 2009. 'The 385+ million word Corpus of Contemporary American English (1990–2008+). Design, architecture, and linguistic insights.' *International Journal of Corpus Linguistics* 14.2: 159–190.
- Davies, Mark. 2010. 'The Corpus of Contemporary American English as the first reliable monitor corpus of English.' *Literary and Linguistic Computing* 25.4: 447–464.
- Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. <<http://corpus2.byu.edu/glowbe/>>.
- Davies, Mark. 2015. 'Why size alone is not enough: The importance of historical, genre-based, and dialectal variation in language.' From data to evidence. Big data, rich data, uncharted data. Helsinki, October 21, 2015.
- Davies, Mark. n.d. *CORPORA: 1.9 billion – 45 million words each: free online access*. <<http://corpus.byu.edu/mutualInformation.asp>> (accessed November 6, 2014).
- Davies, Mark and Robert Fuchs. 2015a. 'A reply.' *English World-Wide* 36.1: 45–47.

Bibliography

- Davies, Mark and Robert Fuchs. 2015b. 'Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE).' *English World-Wide* 36.1: 1–28.
- Desai, Sonalde B., Amaresh Dubey, Brij Lal Joshi, Mitali Sen, Abusaleh Sharif, and Reeve Van-neman. 2010. *Human development in India. Challenges for a society in transition*. Oxford: Oxford University Press.
- Deshors, Sandra C. 2014. 'A case for a unified treatment of EFL and ESL. A multifactorial approach.' *English World-Wide* 35.3: 277–305.
- Deshors, Sandra C. and Stefan Th. Gries. 2014. 'A case for the multifactorial assessment of learner language. The uses of *may* and *can* in French-English interlanguage.' *Corpus methods for semantics. Quantitative studies in polysemy and synonymy*. Ed. Dylan Glynn and Justyna A. Robinson. Amsterdam: Benjamins.
- Deshors, Sandra C. and Stefan Th. Gries. 2016. 'Profiling verb complementation constructions across New Englishes: A two-step random forests analysis of *ing* vs. *to* complements.' *International Journal of Corpus Linguistics* 21.2: 192–218.
- Deterding, David. 2007. *Singapore English*. Edinburgh: Edinburgh University Press.
- Diessel, Holger. 2007. 'Frequency effects in language acquisition, language use, and diachronic change.' *New Ideas in Psychology* 25.2: 108–127.
- Dillard, Joey Lee. 1970. 'Principles in the history of American English—Paradox, virginity, and cafeteria.' *The Florida Foreign Language Reporter* 8: 32–33.
- Dirven, René. 1999. 'Conversion as a conceptual metonymy of event schemata.' *Metonymy in language and thought*. Ed. Klaus-Uwe Panther and Günter Radden. Amsterdam: John Benjamins. 275–287.
- Dixon, Robert M. W. 2005. *A semantic approach to English grammar*. 2nd ed. Oxford: Oxford University Press.
- Don, Jan. 1993. *Morphological conversion*. Utrecht: LEd.
- Don, Jan, Mieke Trommelen, and Wim Zonneveld. 2000. 'Conversion and category indeterminacy.' *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung*. Ed. Geert E. Booij, Christian Lehmann, and Joachim Mugdan. Vol. 1. Berlin: De Gruyter. 943–952.
- Dragon Sea Shipping. 2014. *Company profile*. <<http://en.dragonseas.com/about>> (accessed March 16, 2015).
- Edwards, Alison. 2016. *English in the Netherlands. Functions, forms and attitudes*. Amsterdam: John Benjamins.

- Edwards, Alison and Samantha Laporte. 2015. 'Outer and expanding circle Englishes. The competing roles of norm orientation and proficiency levels.' *English World-Wide* 36.2: 135–169.
- Eickhoff, Carsten and Arjen P. de Vries. 2011. 'How crowdsourcable is your task?' *Proceedings of the workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*. Ed. Matthew Lease, Vitor Carvalho, and Emine Yilmaz. 11–14. <http://ir.ischool.utexas.edu/csdm2011/proceedings/csdm2011_eickhoff.pdf> (accessed April 27, 2015).
- Eickhoff, Carsten and Arjen P. de Vries. 2013. 'Increasing cheat robustness of crowdsourcing tasks.' *Information Retrieval* 16.2: 121–137.
- Ellis, Nick C. and Fernando Ferreira-Junior. 2009. 'Constructions and their acquisition. Islands and the distinctiveness of their occupancy.' *Annual Review of Cognitive Linguistics* 7: 187–200.
- Encyclopædia Britannica Online. 2014. *Continental Divide*. Ed. Encyclopædia Britannica Inc. <<http://www.britannica.com/place/Continental-Divide>> (accessed September 14, 2015).
- Enochson, Kelly and Jennifer Culbertson. 2015. 'Collecting psycholinguistic response time data using Amazon Mechanical Turk.' *PloS one* 10.3: e0116946.
- Evans, Stephen. 2000. 'Hong Kong's new English language policy in education.' *World Englishes* 19.2: 185–204.
- Evans, Stephen. 2009. 'The evolution of the English-language speech community in Hong Kong.' *English World-Wide* 30.3: 278–301.
- Evans, Stephen. 2010. 'Business as usual: The use of English in the professional world in Hong Kong.' *English for Specific Purposes* 29.3: 153–167.
- Evans, Stephen. 2013. 'The long march to biliteracy and trilingualism. Language policy in Hong Kong education since the Handover.' *Annual Review of Applied Linguistics* 33: 302–324.
- Evans, Stephen. 2014. 'The evolutionary dynamics of postcolonial Englishes: A Hong Kong case study.' *Journal of Sociolinguistics* 18.5: 571–603.
- Evans, Stephen. 2015a. 'Modelling the development of English in Hong Kong.' *World Englishes* 34.3: 389–410.
- Evans, Stephen. 2015b. 'Word-formation in Hong Kong English: Diachronic and synchronic perspectives.' *Asian Englishes* 17.2: 116–131.
- Evert, Stefan. 2004. *www.collocations.de – Association measures*. <<http://collocations.de/AM/index.html>> (accessed November 6, 2014).

Bibliography

- Fahrner, Annette. 2016. 'Der Erwerb von *es*-Konstruktionen durch spanischsprachige Deutschlernende.' PhD thesis. Freiburg: Albert-Ludwigs-Universität.
- Farrell, Patrick. 2001. 'Functional shift as category underspecification.' *English Language and Linguistics* 5.1: 109–130.
- Feinberg, Jonathan. 2013. *Wordle. Beautiful word clouds*. <<http://www.wordle.net>> (accessed October 14, 2014).
- Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. Los Angeles: Sage.
- Fletcher, Williams H. 2007. 'Concordancing the web: Promise and problems, tools and techniques.' *Corpus linguistics and the web*. Ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. Amsterdam: Rodopi. 25–45.
- Forster, Kenneth I. 2010. 'Using a maze task to track lexical and sentence processing.' *The Mental Lexicon* 5.3: 347–357.
- Forster, Kenneth I. n.d. *The Word Maze game*. <http://www.u.arizona.edu/~kforster/MAZE/how_it_works.htm> (accessed April 16, 2015).
- Forster, Kenneth I., Christine Guerrero, and Lisa Elliot. 2009. 'The maze task: measuring forced incremental sentence processing time.' *Behavior Research Methods* 41.1: 163–171.
- Free-Press-Release Inc. 2013. *About Free-Press-Release.Com*. <<http://www.free-press-release.com/about-us.html>> (accessed January 13, 2016).
- Frick, Andrea, Marie-Thérèse Bächtiger, and Ulf-Dietrich Reips. 2001. 'Financial incentives, personal information, and drop out in online studies.' *Dimensions of internet science*. Ed. Ulf-Dietrich Reips and Michael Bosnjak. Lengerich: Pabst. 209–219.
- Gargesh, Ravinder. 2006. 'South Asian Englishes.' *The handbook of World Englishes*. Ed. Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson. Malden: Blackwell. 90–113.
- Gargesh, Ravinder. 2008. 'Indian English: phonology.' *Varieties of English. 4. Africa, South and Southeast Asia*. Ed. Rajend Mesthrie. Berlin: Mouton de Gruyter. 231–243.
- Garzone, Giuliana. 2012. 'Where do web genres come from? The case of blogs.' *Evolving genres in web-mediated communication*. Ed. Sandra Campagna, Giuliana Garzone, Cornelia Ilie, and Elizabeth Rowley-Jolivet. Bern: Peter Lang. 217–242.
- Gatto, Maristella. 2014. *The web as corpus. Theory and practice*. London: Bloomsbury.
- Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevelhierarchical models*. Cambridge: Cambridge University Press.
- Gibson, Edward, Steve Piantadosi, and Kristina Fedorenko. 2011. 'Using Mechanical Turk to obtain and analyze English acceptability judgments.' *Language and Linguistics Compass* 5.8: 509–524.

- Gilquin, Gaëtanelle. 2015. 'At the interface of contact linguistics and second language acquisition research. New Englishes and learner Englishes compared.' *English World-Wide* 36.1: 91–124.
- Gilquin, Gaëtanelle and Sylviane Granger. 2011. 'From EFL to ESL. Evidence from the *International Corpus of Learner English*.' *Exploring second-language varieties of English and learner Englishes. Bridging a paradigm gap*. Ed. Joybrato Mukherjee and Marianne Hundt. Amsterdam: John Benjamins. 55–78.
- Gilquin, Gaëtanelle and Stefan Th. Gries. 2009. 'Corpora and experimental methods: A state-of-the-art review.' *Corpus Linguistics and Linguistic Theory* 5.1: 1–26.
- Gisborne, Nikolas. 2009. 'Aspects of the morphosyntactic typology of Hong Kong English.' *English World-Wide* 30.2: 149–169.
- Goh, Robbie B. H. 2013. 'Uncertain locale. The dialectics of space and the cultural politics of English in Singapore.' *The politics of English. South Asia, Southeast Asia and the Asia Pacific*. Ed. Lionel Wee, Robbie B. H. Goh, and Lisa Lim. Amsterdam: John Benjamins. 125–143.
- Goldberg, Adele E. 1995. *Constructions. A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.
- Goldberg, Adele E. 2006. *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.
- Goldberg, Adele E. 2011. 'Corpus evidence of the viability of statistical preemption.' *Cognitive Linguistics* 22.1: 131–153.
- Görlach, Manfred. 1997. 'Celtic Englishes?' *The Celtic Englishes*. Ed. Hildegard L. C. Tristram. Heidelberg: Winter. 27–54.
- Görlach, Manfred. 2002. *Still more Englishes*. Amsterdam: John Benjamins.
- Grace-Martin, Karen. 2014. *When NOT to center a predictor variable in regression*. <<http://www.theanalysisfactor.com/when-not-to-center-a-predictor-variable-in-regression/>> (accessed December 4, 2014).
- Grace-Martin, Karen. n.d. *Can Likert scale data ever be continuous?* <<http://www.theanalysisfactor.com/can-likert-scale-data-ever-be-continuous/>> (accessed January 13, 2016).
- Greenbaum, Sidney. 1996. 'Introducing ICE.' *Comparing English worldwide. The International Corpus of English*. Ed. Sidney Greenbaum. Oxford: Clarendon Press. 3–12.
- Greenbaum, Sidney and Gerald Nelson. 1996. 'The International Corpus of English (ICE) Project.' *World Englishes* 15.1: 3–15.
- Grice, H. Paul. 1975. 'Logic and conversation.' *Syntax and semantics. Vol. 3, Speech acts*. Ed. Peter Cole and Jerry L. Morgan. New York: Academic Press. 41–58.

Bibliography

- Gries, Stefan Th. 2002. 'Evidence in linguistics: Three approaches to genitives in English.' *LACUS Forum XXVIII: What constitutes evidence in linguistics?* Ed. Ruth M. Brend, William J. Sullivan, and Arle R. Lommel. Fullerton: LACUS. 17–31.
- Gries, Stefan Th. 2009. *Statistics for linguistics with R. A practical introduction*. Berlin: Mouton de Gruyter.
- Gries, Stefan Th. 2013a. '50-something years of work on collocations. What is or should be next...' *International Journal of Corpus Linguistics* 18.1: 137–165.
- Gries, Stefan Th. 2013b. *Statistics for linguistics with R. A practical introduction*. 2nd ed. Berlin: De Gruyter Mouton.
- Gries, Stefan Th. 2015. 'The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models.' *Corpora* 10.1: 95–125.
- Gries, Stefan Th., Beate Hampe, and Doris Schönefeld. 2005. 'Converging evidence. Bringing together experimental and corpus data on the association of verbs and constructions.' *Cognitive Linguistics* 16.4.
- Gries, Stefan Th. and Joybrato Mukherjee. 2010. 'Lexical gravity across varieties of English. An ICE-based study of n-grams in Asian Englishes.' *International Journal of Corpus Linguistics* 15.4: 520–548.
- Grieve, Jack, Douglas Biber, Eric Friginal, and Tatiana Nekrasova. 2010. 'Variation among blogs: A multi-dimensional analysis.' *Genres on the web. Computational models and empirical studies*. Ed. Alexander Mehler, Serge Sharoff, and Marina Santini. Dordrecht: Springer. 303–321. <<http://dl.dropbox.com/u/99161057/GrieveetalBlogs.pdf>> (accessed January 22, 2015).
- Groves, Julie M. 2011. "'Linguistic schizophrenia' in Hong Kong.' *English Today* 27.4: 33–42.
- Gut, Ulrike. 2004. 'Nigerian English: phonology.' *A handbook of varieties of English. A multimedia reference tool; two volumes*. Ed. Bernd Kortmann and Edgar W. Schneider. Berlin: Mouton de Gruyter. 813–830.
- Gut, Ulrike. 2007. 'First language influence and final consonant clusters in the New Englishes of Singapore and Nigeria.' *World Englishes* 26.3: 346–359.
- Gut, Ulrike. 2009. 'Past tense marking in Singapore English verbs.' *English World-Wide* 30.3: 262–277.
- Hansen, Beke. 2015. 'The ICE metadata: A window to the past? An exploratory study of Hong Kong English.' From data to evidence. Big data, rich data, uncharted data. Helsinki, October 20, 2015.
- Harrell, Frank E. 2015. *Regression Modeling Strategies*. 2nd ed. Cham: Springer.

- Haselow, Alexander. 2010. 'Thomas Biermeier, Word-formation in New Englishes: A corpus-based analysis.' *Anglia* 128.1: 131–135.
- Haselow, Alexander. 2011. *Typological changes in the lexicon. Analytic tendencies in English noun formation*. Berlin: De Gruyter.
- Hashim, Azirah and Gerhard Leitner. 2011. 'Contact expressions in contemporary Malaysian English.' *World Englishes* 30.4: 551–568.
- Hasselgren, Angela. 1994. 'Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary.' *International Journal of Applied Linguistics* 4.2: 237–260.
- Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
- Heller, Benedikt and Melanie Röthlisberger. 2015. 'Big data on trial. Researching syntactic alternations in GloWbE and ICE.' From data to evidence. Big data, rich data, uncharted data. Helsinki, October 21, 2015.
- Hergenhahn, B. R. and Tracy B. Henley. 2014. *An introduction to the history of psychology*. 7th ed. Belmont: Wadsworth Cengage Learning.
- Hilpert, Martin. 2014a. *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.
- Hilpert, Martin. 2014b. *Martin Hilpert's motion chart resource page*. <<http://members.unine.ch/martin.hilpert/motion.html>> (accessed October 7, 2014).
- Hoffmann, Sebastian. 2009. 'Corpus linguistics and the internet – An overview and three case studies.' *Anglistik: International Journal of English Studies* 20.1: 23–39.
- Hoffmann, Sebastian, Marianne Hundt, and Joybrato Mukherjee. 2011. 'Indian English – An emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes.' *Anglia* 129.3-4: 258–280.
- Hoffmann, Sebastian, Andrea Sand, and Peter K.W. Tan. 2012. 'The Corpus of Historical Singapore English. A first pilot study of data from the 1950s and 1960s.' ICAME 33. Leuven. <http://www.ling.arts.kuleuven.be/icame33/_pdf/icame33abstracts.pdf>.
- Hoffmann, Thomas. 2014. 'The cognitive evolution of Englishes. The role of constructions in the Dynamic Model.' *The evolution of Englishes. The Dynamic Model and beyond*. Ed. Sarah Buschfeld, Thomas Hoffmann, Magnus Huber, and Alexander Kautzsch. Amsterdam: John Benjamins. 160–180.
- Hoffmann, Thomas. 2015. 'Constructions and the Dynamic Model: Comparative correlative constructions in Englishes around the world.' ICAME 36. Trier, May 28, 2015.
- Horch, Clemens. 2015. *QualityCrowd2*. <<https://github.com/clorch/QualityCrowd2>>.

Bibliography

- Hoßfeld, Tobias, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. 2013. *CrowdTesting: A novel methodology for subjective user studies and QoE evaluation*. <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.391.4386&rep=rep1&type=pdf>> (accessed April 27, 2015).
- Howe, Jeff. 2006. 'The rise of crowdsourcing.' *Wired* 14.6. <http://archive.wired.com/wired/archive/14.06/crowds.html?pg=1&topic=crowds&topic_set=> (accessed April 28, 2015).
- Hundt, Marianne and Christian Mair. 1999. "Agile" and "uptight" genres: The corpus-based approach to language change in progress.' *International Journal of Corpus Linguistics* 4.2: 221–242.
- Hundt, Marianne, Nadja Nesselhauf, and Carolin Biewer, eds. 2007. *Corpus linguistics and the web*. Amsterdam: Rodopi.
- Hung, Tony T. N. 2012. 'Hong Kong English.' *English in Southeast Asia. Features, policy and language in use*. Ed. Ee-Ling Low and Azirah Hashim. Amsterdam: John Benjamins. 113–133.
- Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Janssen, Niels and Horacio A. Barber. 2012. 'Phrase frequency effects in language production.' *PLoS ONE* 7.3: e33202.
- Jefferson, Gail. 2004. 'Glossary of transcript symbols with an introduction.' *Conversation analysis. Studies from the first generation*. Ed. Gene H. Lerner. Amsterdam: John Benjamins. 13–31. <http://www.liso.ucsb.edu/liso_archives/Jefferson/Transcript.pdf> (accessed January 29, 2015).
- Jiang, Nan. 2012. *Conducting reaction time research in second language studies*. New York: Routledge.
- Johnson, Robert Keith. 1994. 'Language policy and planning in Hong Kong.' *Annual Review of Applied Linguistics* 1993/1994.14: 177–199.
- Kachru, Braj B. 1985. 'Standards, codification and sociolinguistic realism: The English language in the Outer Circle.' *English in the world: Teaching and learning the language and literatures*. Ed. Randolph Quirk and Henry G. Widdowson. Cambridge: Cambridge University Press. 11–30.
- Kachru, Braj B. 1994. 'English in South Asia.' *The Cambridge history of the English Language*. Ed. Robert Burchfield. Cambridge: Cambridge University Press. 497–553.
- Kachru, Yamuna. 2006. *Hindi*. Amsterdam: John Benjamins.
- Kailuweit, Rolf. 2009. 'Konzeptionelle Mündlichkeit!? Überlegungen zur Chat-Kommunikation anhand französischer, italienischer und spanischer Materialien.' *Philologie im Netz* 48: 1–19. <<http://www.phin.de/phin48/p48i.htm>> (accessed July 16, 2014).

- Karius, Ilse. 1985. *Die Ableitung der denominalen Verben mit Nullsuffigierung im Englischen*. Tübingen: Niemeyer.
- Kastovsky, Dieter. 1982. *Wortbildung und Semantik*. Düsseldorf: Schwann-Bagel.
- Keimel, Christian, Julian Habigt, Clemens Horch, and Klaus Diepold. 2012. 'QualityCrowd – A framework for crowd-based quality evaluation.' *Proceedings, 2012 Picture Coding Symposium*. Ed. Marek Domański, Tomasz Grajek, Damian Karwowski, and Ryszard Stasiński. Piscataway: IEEE. 245–248.
- Kirkpatrick, Andy. 2012. 'Theoretical issues.' *English in Southeast Asia. Features, policy and language in use*. Ed. Ee-Ling Low and Azirah Hashim. Amsterdam: John Benjamins. 13–31.
- Kirkpatrick, Andy and Andrew Moody. 2009. 'A tale of two songs: Singapore versus Hong Kong.' *ELT Journal* 63.3: 265–271.
- Knowles, Gerry and Zuraidah Mohd Don. 2003. 'Tagging a corpus of Malay texts, and coping with 'syntactic drift'.' *UCREL Technical Paper number 16. Special issue. Proceedings of the Corpus Linguistics 2003 conference*. Ed. Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery. 422–428. <<http://ucrel.lancs.ac.uk/publications/CL2003/papers/knowles.pdf>> (accessed September 3, 2015).
- Koch, Peter and Wulf Oesterreicher. 1985. 'Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte.' *Romanistisches Jahrbuch* 36: 15–43.
- Koch, Peter and Wulf Oesterreicher. 2007. *Lengua hablada en la Romania: Español, francés, italiano*. Madrid: Gredos.
- Kortmann, Bernd and Kerstin Lunkenheimer. 2013a. 'Introduction.' *The Electronic World Atlas of Varieties of English*. Ed. Bernd Kortmann and Kerstin Lunkenheimer. Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://ewave-atlas.org/introduction>> (accessed March 11, 2015).
- Kortmann, Bernd and Kerstin Lunkenheimer, eds. 2013b. *The Electronic World Atlas of Varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://ewave-atlas.org>>.
- Kortmann, Bernd and Benedikt Szmrecsanyi. 2009. 'World Englishes between simplification and complexification.' *World Englishes – Problems, properties and prospects. Selected papers from the 13th IAWE conference*. Ed. Thomas Hoffmann and Lucia Siebers. Amsterdam: John Benjamins. 265–285. <http://www.benszm.net/omnibuslit/KoSz_pageproofs_IAWE_modified.pdf> (accessed January 9, 2013).

Bibliography

- Kosinski, Robert J. 2013. *A literature review on reaction time*. <<http://biae.clemson.edu/bpc/bp/lab/110/reaction.htm>> (accessed May 7, 2015).
- Koziol, Herbert. 1937. *Handbuch der englischen Wortbildungslehre*. Heidelberg: Winter.
- Krantz, John H. and Reeshad Dalal. 2000. 'Validity of web-based psychological research.' *Psychological experiments on the internet*. Ed. Michael H. Birnbaum. San Diego: Academic Press. 35–60.
- Krug, Manfred. 1998. 'British English is developing a new discourse marker, *innit?* A study in lexicalisation based on social, regional and stylistic variation.' *Arbeiten aus Anglistik und Amerikanistik* 23.2: 145–197.
- kryzk. 2014. *Imagines. Urban Dictionary*. Ed. Urban Dictionary LLC. <<http://www.urbandictionary.com/define.php?term=Imagine>> (accessed March 6, 2015).
- Kunter, Gero. 2015. 'Effects of processing complexity in perception and production. The case of English comparative alternation.' *Word structure and word usage. Proceedings of the NETWords Final Conference*. Ed. Vito Pirrelli, Claudia Marzi, and Marcello Ferro. Vol. 1347. <http://ceur-ws.org/>. 32–36. <<http://ceur-ws.org/Vol-1347/paper07.pdf>> (accessed June 2, 2015).
- Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. 'Age-of-acquisition ratings for 30,000 English words.' *Behavior Research Methods* 44.4: 978–990.
- Labov, William. 2001. *Principles of linguistic change. Vol. 2: Social factors*. Chichester: Blackwell.
- Lai, Mee-Ling. 2005. 'Language attitudes of the first postcolonial generation in Hong Kong secondary schools.' *Language in Society* 34.3: 363–388.
- Lai, Mee-Ling. 2012. 'Tracking language attitudes in postcolonial Hong Kong. An interplay of localization, mainlandization, and internationalization.' *Multilingua* 31: 83–111.
- Langacker, Ronald W. 1987. 'Nouns and verbs.' *Language* 63.1: 53–94.
- Lange, Claudia. 2012. *The syntax of spoken Indian English*. Amsterdam: John Benjamins.
- Laporte, Samantha. 2012. 'Mind the gap! Bridge between World Englishes and learner Englishes in the making.' *English Text Construction* 5.2: 264–291.
- Lauwers, Peter. 2008. 'The nominalization of adjectives in French: From morphological conversion to categorial mismatch.' *Folia Linguistica* 42.1: 135–176.
- Leech, Geoffrey. 2007. 'New resources, or just better old ones? The Holy Grail of representativeness.' *Corpus linguistics and the web*. Ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. Amsterdam: Rodopi. 133–149.
- Leimgruber, Jakob R. E. 2013a. *Singapore English. Structure, variation, and usage*. Cambridge: Cambridge University Press.

- Leimgruber, Jakob R. E. 2013b. 'The management of multilingualism in a city-state: Language policy in Singapore.' *Multilingualism and language diversity in urban areas. Acquisition, identities, space, education*. Ed. Peter Siemund, Ingrid Gogolin, Monika Edith Schulz, and Julia Davydova. Amsterdam: John Benjamins. 227–256.
- Leimgruber, Jakob R. E. 2013c. 'The trouble with World Englishes.' *English Today* 29.3: 3–7.
- Leisi, Ernst and Christian Mair. 2008. *Das heutige Englisch. Wesenszüge und Probleme*. 9th ed. Heidelberg: Winter.
- Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins.
- Li, David C. S. 2000. 'Cantonese-English code-switching research in Hong Kong: A Y2K review.' *World Englishes* 19.3: 305–322.
- Lim, Lisa, ed. 2004. *Singapore English. A grammatical description*. Amsterdam: John Benjamins.
- Lim, Lisa, Anne Pakir, and Lionel Wee. 2010. 'English in Singapore: Policies and prospects.' *English in Singapore. Modernity and management*. Ed. Lisa Lim, Anne Pakir, and Lionel Wee. Hong Kong: Hong Kong University Press. 3–18.
- Lorenz, David. 2013. *Contractions of English semi-modals*. Freiburg: Rombach.
- Low, Ee-Ling. 2012. 'Singapore English.' *English in Southeast Asia. Features, policy and language in use*. Ed. Ee-Ling Low and Azirah Hashim. Amsterdam: John Benjamins. 35–53.
- Low, Ee-Ling and Azirah Hashim, eds. 2012. *English in Southeast Asia. Features, policy and language in use*. Amsterdam: John Benjamins.
- Luke, Kang-Kwong and Jack C. Richards. 1982. 'English in Hong Kong: Functions and status.' *English World-Wide* 3.1: 47–64.
- Mair, Christian. 2007. 'Change and variation in present-day English: Integrating the analysis of closed corpora and web-based monitoring.' *Corpus linguistics and the web*. Ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. Amsterdam: Rodopi. 233–247.
- Mair, Christian. 2013a. 'Speculating on the future of English as a contact language.' *English as a contact language*. Ed. Daniel Schreier and Marianne Hundt. Cambridge: Cambridge University Press. 314–328.
- Mair, Christian. 2013b. 'The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars.' *English World-Wide* 34.3: 253–278.
- Mair, Christian. 2015. 'Response to Davies and Fuchs.' *English World-Wide* 36.1: 29–33.
- Marchand, Hans. 1960. *The categories and types of present-day English word-formation. A synchronic-diachronic approach*. Wiesbaden: Otto Harrassowitz.

Bibliography

- Marchand, Hans. 1969. *The categories and types of present-day English word-formation. A synchronic-diachronic approach*. 2nd ed. München: C. H. Beck.
- Mason, Winter and Siddharth Suri. 2012. 'Conducting behavioral research on Amazon's Mechanical Turk.' *Behavior Research Methods* 44.1: 1–23.
- Matthews, Stephen and Virginia Yip. 1994. *Cantonese. A comprehensive grammar*. London: Routledge.
- Maxwell, William Edward. 1907. *A manual of the Malay language with an introductory sketch of the Sanskrit element in Malay*. 8th ed. London: Kegan Paul, Trench, Trübner, & Co. <<http://www.gutenberg.org/files/25604/25604-h/25604-h.htm>> (accessed September 3, 2015).
- McAllister Byun, Tara, Peter F. Halpin, and Daniel Szeredi. 2015. 'Online crowdsourcing for efficient rating of speech: A validation study.' *Journal of Communication Disorders* 53: 70–83.
- McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics. Method, theory and practice*. Cambridge: Cambridge University Press.
- McGraw, Kenneth O., Mark D. Tew, and John E. Williams. 2000. 'The integrity of web-delivered experiments: Can you trust the data?' *Psychological Science* 11.6: 502–506.
- Mesthrie, Rajend. 2008. 'Synopsis: morphological and syntactic variation in Africa and South and Southeast Asia.' *Varieties of English. 4. Africa, South and Southeast Asia*. Ed. Rajend Mesthrie. Berlin: Mouton de Gruyter. 624–635.
- Mesthrie, Rajend and Rakesh M. Bhatt. 2008. *World Englishes. The study of new linguistic varieties*. Cambridge: Cambridge University Press. <<http://www.loc.gov/catdir/enhancements/fy0808/2008003210-b.html>>.
- Metin, Senem Kumova and Bahar Karaoğlan. 2011. 'Measuring collocation tendency of words.' *Journal of Quantitative Linguistics* 18.2: 174–187.
- Meunier, Fanny and Damien Littré. 2013. 'Tracking Learners' Progress. Adopting a Dual 'Corpus cum Experimental Data' Approach.' *The Modern Language Journal* 97.S1: 61–76.
- Michaelis, Laura. 2004. 'Type shifting in Construction Grammar: An integrated approach to aspectual coercion.' *Cognitive Linguistics* 15.1: 1–67.
- Milroy, Lesley. 1980. *Language and social networks*. Oxford: Basil Blackwell.
- Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics. Method and interpretation*. Malden: Blackwell.
- Mufwene, Salikoko S. 1990. 'Transfer and the substrate hypothesis in creolistics.' *Studies in Second Language Acquisition* 12.1: 1–23.

- Mukherjee, Joybrato. 2007. 'Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium.' *Journal of English Linguistics* 35.2: 157–187.
- Mukherjee, Joybrato. 2009a. *Anglistische Korpuslinguistik. Eine Einführung*. Berlin: Schmidt.
- Mukherjee, Joybrato. 2009b. 'The lexicogrammar of present-day Indian English. Corpus-based perspectives on structural nativisation.' *Exploring the lexis-grammar interface*. Ed. Ute Römer and Rainer Schulze. Amsterdam: John Benjamins. 117–135.
- Mukherjee, Joybrato and Stefan Th. Gries. 2009. 'Collostructional nativisation in New Englishes. Verb-construction associations in the International Corpus of English.' *English World-Wide* 30.1: 27–51.
- Mukherjee, Joybrato and Marianne Hundt, eds. 2011. *Exploring second-language varieties of English and learner Englishes. Bridging a paradigm gap*. Amsterdam: John Benjamins.
- Mukherjee, Joybrato and Marco Schilk. 2012. 'Exploring variation and change in New Englishes. Looking into the International Corpus of English (ICE) and beyond.' *The Oxford handbook of the history of English*. Ed. Terttu Nevalainen and Elizabeth Closs Traugott. Oxford: Oxford University Press. 189–199.
- Musch, Jochen and Ulf-Dietrich Reips. 2000. 'A brief history of web experimenting.' *Psychological experiments on the internet*. Ed. Michael H. Birnbaum. San Diego: Academic Press. 61–87.
- Myers, Scott. 1984. 'Zero-derivation and inflection.' *MIT Working Papers in Linguistics* 7: 53–69.
- Nelson, Gerald. 2002. *Markup manual for spoken texts*. <<http://ice-corpora.net/ice/spoken.doc>> (accessed February 3, 2015).
- Nelson, Gerald. 2004. 'Introduction.' *World Englishes* 23.2: 225–226.
- OED Online*. n.d. Oxford University Press.
- Onysko, Alexander. 2016. 'Modeling world Englishes from the perspective of language contact.' *World Englishes* 35.2: 196–220.
- Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. 'Instructional manipulation checks. Detecting satisficing to increase statistical power.' *Journal of Experimental Social Psychology* 45.4: 867–872.
- Osborne, Jason W. and Amy Overbay. 2004. 'The power of outliers (and why researchers should ALWAYS check for them).' *Practical Assessment, Research & Evaluation* 9.6. <<http://pareonline.net/getvn.asp?v=9&n=6>> (accessed November 24, 2015).
- Pang, Terence T. T. 2003. 'Hong Kong English: A stillborn variety?' *English Today* 19.2: 12–18.
- Pavesi, Maria. 1998. "Same word, same idea". Conversion as a word formation process.' *International Review of Applied Linguistics in Language Teaching* 36.3: 213–231.

Bibliography

- Percillier, Michael. 2016. *World Englishes and second language acquisition. Insights from South-east Asian Englishes*. Amsterdam: John Benjamins Publishing Company.
- Perek, Florent and Martin Hilpert. 2014. 'Constructional tolerance. Cross-linguistic differences in the acceptability of non-conventional uses of constructions.' *Constructions and Frames* 6.2: 266–304.
- Pinheiro, José C. and Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Plag, Ingo. 1999. *Morphological productivity. Structural constraints in English derivation*. Berlin: Mouton de Gruyter.
- Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.
- Platt, John Talbot, Heidi Weber, and Mian Lian Ho. 1983. *Singapore and Malaysia*. Amsterdam: John Benjamins.
- Pliatsikas, Christos, Linda Wheeldon, Aditi Lahiri, and Peter C. Hansen. 2014. 'Processing of zero-derived words in English: An fMRI investigation.' *Neuropsychologia* 53: 47–53.
- Po-Ching, Yip and Don Rimmington. 2004. *Chinese. A comprehensive grammar*. London: Routledge.
- Prince, Alan and Paul Smolensky. 2004. *Optimality theory. Constraint interaction in Generative Grammar*. Oxford: Blackwell.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1972. *A grammar of contemporary English*. London: Longman.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
- R Core Team. 2014. *R: A language and environment for statistical computing*. Wien. <<http://www.R-project.org>>.
- Rainer, Franz. 1988. 'Towards a theory of blocking: The case of Italian and German quality nouns.' *Yearbook of Morphology 1988*. Ed. Geert E. Booij and Jaap van Marle. Dordrecht: Foris. 155–185.
- Rajeshodayanchal at Malayalam Wikipedia. 2011. *malayaaLam: inthya yuTe bhupaTam [Map of India]*. Ed. Wikipedia, The Free Encyclopedia. <<https://commons.wikimedia.org/wiki/File:India-map-en.svg>> (accessed August 9, 2015).
- Ratcliff, Roger, Anjali Thapar, Pablo Gomez, and Gail McKoon. 2004. 'A diffusion model analysis of the effects of aging in the lexical-decision task.' *Psychology and Aging* 19.2: 278–289.
- Reips, Ulf-Dietrich. 2002. 'Standards for internet-based experimenting.' *Experimental Psychology* 49.4: 243–256.

- Rohdenburg, Günter. 1996. 'Cognitive complexity and increased grammatical explicitness in English.' *Cognitive Linguistics* 7.2: 149–182.
- van Rooy, Bertus. 2011. 'A principled distinction between error and conventionalized innovation in African Englishes.' *Exploring second-language varieties of English and learner Englishes. Bridging a paradigm gap*. Ed. Joybrato Mukherjee and Marianne Hundt. Amsterdam: John Benjamins. 189–207.
- Ross, Claudia and Jing-heng Sheng Ma. 2006. *Modern Mandarin Chinese grammar. A practical guide*. London, New York: Routledge.
- Rowley-Jolivet, Elizabeth. 2012. 'Open science and the re-purposing of genre: An analysis of web-mediated laboratory protocols.' *Evolving genres in web-mediated communication*. Ed. Sandra Campagna, Giuliana Garzone, Cornelia Ilie, and Elizabeth Rowley-Jolivet. Bern: Peter Lang. 127–149.
- RStudio. n.d. *RStudio: Integrated development environment for R*. Boston. <<http://www.rstudio.com>>.
- Sailaja, Pingali. 2009. *Indian English*. Edinburgh: Edinburgh University Press.
- Säily, Tanja and Jukka Suomela. 2009. 'Comparing type counts: The case of women, men and -ity in early English letters.' *Corpus linguistics*. Ed. Antoinette Renouf and Andrew Kehoe. Vol. 69. Language and computers. Amsterdam: Rodopi. 87–109.
- Sanders, Gerald. 1988. 'Zero derivation and the overt analogue criterion.' *Theoretical morphology. Approaches in modern linguistics*. Ed. Michael Hammond and Noonan Michael. San Diego: Academic Press. 155–175.
- Santini, Marina, Alexander Mehler, and Serge Sharoff. 2010. 'Riding the rough waves of genre on the web. Concepts and research questions.' *Genres on the web. Computational models and empirical studies*. Ed. Alexander Mehler, Serge Sharoff, and Marina Santini. Vol. 42. Dordrecht: Springer. 3–30.
- Sayers, Dave. 2014. 'The mediated innovation model: A framework for researching media influence in language change.' *Journal of Sociolinguistics* 18.2: 185–212.
- Schegloff, Emanuel A., Gail Jefferson, and Harvey Sacks. 1977. 'The preference for self-correction in the organization of repair in conversation.' *Language* 53.2: 361–382.
- Schilk, Marco. 2011. *Structural nativization in Indian English lexicogrammar*. Amsterdam: John Benjamins.
- Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells. From corpus to cognition*. Berlin: Mouton de Gruyter.

Bibliography

- Schmid, Hans-Jörg. 2007. 'Entrenchment, salience, and basic levels.' *The Oxford handbook of Cognitive Linguistics*. Ed. Dirk Geeraerts and Hubert Cuyckens. Oxford: Oxford University Press. 117–138.
- Schmid, Hans-Jörg. 2010. 'Does frequency in text instantiate entrenchment in the cognitive system?' *Quantitative methods in cognitive semantics. Corpus-driven approaches*. Ed. Dylan Glynn and Kerstin Fischer. Berlin: De Gruyter. 101–133.
- Schmid, Hans-Jörg. 2011. *English morphology and word-formation. An introduction*. 2nd ed. Berlin: Schmidt.
- Schmid, Hans-Jörg. 2015. 'A blueprint of the Entrenchment-and-Conventionalization Model.' *Yearbook of the German Cognitive Linguistics Association 3*. Ed. Peter Uhrig and Thomas Herbst. Vol. 3. Berlin: De Gruyter. 3–25.
- Schneider, Edgar W. 2003. 'The dynamics of New Englishes. From identity construction to dialect birth.' *Language* 79.2: 233–281.
- Schneider, Edgar W. 2007. *Postcolonial English. Varieties around the world*. Cambridge: Cambridge University Press.
- Schneider, Edgar W. 2008. 'Accommodation versus identity? A response to Trudgill.' *Language in Society* 37.2: 262–267.
- Schneider, Edgar W. 2012. 'Exploring the interface between World Englishes and second language acquisition – and implications for English as a lingua franca.' *Journal of English as a Lingua Franca* 1.1: 57–91.
- Schneider, Edgar W. 2014a. 'Asian Englishes – into the future: a bird's eye view.' *Asian Englishes* 16.3: 249–256.
- Schneider, Edgar W. 2014b. 'New reflections on the evolutionary dynamics of World Englishes.' *World Englishes* 33.1: 9–32.
- Schneider, Ulrike. 2014c. *Frequency, hesitation and chunks. A usage-based study of chunking in English*. Freiburg: Rombach.
- Schnell, Rainer, Paul B. Hill, and Elke Esser. 2011. *Methoden der empirischen Sozialforschung*. 9th ed. München: Oldenbourg.
- Schnoebelen, Tyler and Victor Kuperman. 2010. 'Using Amazon Mechanical Turk for linguistic research.' *Psihologija* 43.4: 441–464.
- Schnurr, Stephanie and Angela Chan. 2009. 'Politeness and leadership discourse in New Zealand and Hong Kong: A cross-cultural case study of workplace talk.' *Journal of Politeness Research* 5.2: 131–157.
- Schönefeld, Doris. 2011. 'Introduction. On evidence and the convergence of evidence in linguistic research.' *Converging evidence. Methodological and theoretical issues for linguistic*

- research. Ed. Doris Schönefeld. Vol. 33. Human cognitive processing. Amsterdam: John Benjamins. 1–31.
- Schütze, Carson T. 1996. *The empirical base of linguistics. Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
- Schütze, Carson T. and Jon Sprouse. 2013. 'Judgment data.' *Research methods in linguistics*. Ed. Robert J. Podesva and Devyani Sharma. Cambridge: Cambridge University Press. 27–50.
- Schwarz, Gideon. 1978. 'Estimating the Dimension of a Model.' *The Annals of Statistics* 6.2: 461–464.
- Sedlatschek, Andreas. 2009. *Contemporary Indian English. Variation and change*. Amsterdam: John Benjamins.
- Seoane, Elena and Cristina Suárez-Gómez. 2013. 'The expression of the perfect in East and South-East Asian Englishes.' *English World-Wide* 34.1: 1–25.
- Setter, Jane, Cathy S. P. Wong, and Brian Hok-Shing Chan. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.
- Shi, Dignxu. 2006. 'Hong Kong written Chinese. Language change induced by language contact.' *Journal of Asian Pacific Communication* 16.2: 299–318.
- Siegel, Jeff. 1999. 'Transfer constraints and substrate influence in Melanesian Pidgin.' *Journal of Pidgin and Creole Languages* 14.1: 1–44.
- Siemund, Peter, Monika Edith Schulz, and Martin Schweinberger. 2014. 'Studying the linguistic ecology of Singapore: A comparison of college and university students.' *World Englishes* 33.3: 340–362.
- Siew Imm, Tan. 2009. 'Lexical borrowing in Malaysian English: Influences of Malay.' *Lexis – E-Journal in English Lexicology* 3 (L'emprunt/Borrowing): 11–62. <http://lexis.univ-lyon3.fr/IMG/pdf/Lexis_3_Imm.pdf> (accessed September 7, 2015).
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.' *Psychological Science* 22.11: 1359–1366.
- Siyanova, Anna and Norbert Schmitt. 2008. 'L2 learner production and processing of collocation. A multi-study perspective.' *The Canadian Modern Language Review* 64.3: 429–458.
- Slinker, Bryan K. and Stanton A. Glantz. 1988. 'Multiple linear regression is a useful alternative to traditional analyses of variance.' *The American Journal of Physiology* 255.3: R353–R367.
- Sridhar, Kamal K. and S. N. Sridhar. 1986. 'Bridging the paradigm gap: Second language acquisition theory and indigenized varieties of English.' *World Englishes* 5.1: 3–14.

Bibliography

- Sridhar, S. N. 1988. 'Language variation, attitudes, and rivalry. The spread of Hindi in India.' *Language spread and language policy*. Ed. Peter H. Lowenberg. Washington, D.C.: Georgetown University Press. 300–319.
- Srivastava, Ravindra Nath. 1994. *Studies in language and linguistics. Vol. 4. Applied linguistics*. Delhi: Kalinga.
- StarCraft Wiki. n.d. *Fast expand*. Ed. StarCraft Wiki. <http://starcraft.wikia.com/wiki/Fast_expand> (accessed March 9, 2015).
- Štekauer, Pavol, Salvador Valera, and Livia Körtvélyessy. 2012. *Word-formation in the world's languages. A typological survey*. Cambridge: Cambridge University Press.
- Strang, Barbara M. H. 1970. *A history of English*. London: Routledge.
- Stubbs, Michael. 1995. 'Collocations and semantic profiles. On the cause of the trouble with quantitative studies.' *Functions of Language* 2.1: 23–55.
- Suttle, Laura and Adele E. Goldberg. 2011. 'The partial productivity of constructions as induction.' *Linguistics* 49.6.
- Sweet, Henry. 1891. *A new English grammar. Logical and historical*. Oxford: Clarendon Press.
- Szmrecsanyi, Benedikt. 2009. 'Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English.' *Language Variation and Change* 21.3: 319–353.
- Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
- Tagliamonte, Sali A. 2014. 'Situating media influence in sociolinguistic context.' *Journal of Sociolinguistics* 18.2: 223–232.
- Tan, Ying-Ying. 2014. 'English as a 'mother tongue' in Singapore.' *World Englishes* 33.3: 319–339.
- Teddiman, Laura. 2012. 'Conversion and the lexicon: Comparing evidence from corpora and experimentation.' *Frequency effects in language representation*. Ed. Dagmar Divjak and Stefan Th. Gries. Vol. 244.2. Berlin: De Gruyter. 235–254.
- Tent, Jan. 2001. 'A profile of the Fiji English lexis.' *English World-Wide* 22.2: 209–245.
- Terassa, Laura. in preparation. 'Competing factors in simplification: Frequency, substratum transfer, and institutionalization.' PhD thesis. Freiburg: Albert-Ludwigs-Universität.
- The ICE Project. 2009. *Corpus design*. <<http://ice-corpora.net/ICE/design.htm>> (accessed November 28, 2014).
- The ICE Project. 2015. *International Corpus of English*. <<http://ice-corpora.net/ice/index.htm>> (accessed January 4, 2016).

- Thomason, Sarah Grey and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.
- Traugott, Elizabeth Closs. 2007. 'The concepts of constructional mismatch and type-shifting from the perspective of grammaticalization.' *Cognitive Linguistics* 18.4: 523–557.
- Traugott, Elizabeth Closs and Graeme Trousdale. 2013. *Constructionalization and constructional changes*. Oxford: Oxford University Press.
- Trudgill, Peter. 2004. *New-dialect formation the inevitability of colonial Englishes*. Oxford [u.a.]: Oxford Univ. Press.
- Trudgill, Peter. 2008. 'Colonial dialect contact in the history of European languages. On the irrelevance of identity to new-dialect formation.' *Language in Society* 37.2: 241–254.
- Tschichold, Cornelia. 2002. 'Learner English.' *Perspectives on English as a world language*. Ed. D. J. Allerton, Paul Skandera, and Cornelia Tschichold. Basel: Schwabe. 125–133.
- Ungerer, Friedrich. 2002. 'The conceptual function of derivational word-formation in English.' *Anglia* 120.4.
- University Centre for Computer Corpus Research on Language. 2015. *CLAWS part-of-speech tagger for English*. <<http://ucrel.lancs.ac.uk/claws/>> (accessed March 25, 2015).
- Urban Dictionary LLC. 1999. *Urban Dictionary*. Ed. Urban Dictionary LLC. San Francisco. <<http://www.urbandictionary.com>> (accessed March 6, 2015).
- Vaish, Viniti. 2008. 'Mother tongues, English, and religion in Singapore.' *World Englishes* 27.3/4: 450–464.
- Váradi, Tamás. 2001. 'The linguistic relevance of corpus linguistics.' *Proceedings of the Corpus Linguistics 2001 Conference*. Ed. Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja. Vol. 13. UCREL Technical Papers. Lancaster: UCREL. 587–593.
- Wald, Benji. 1993. 'On the evolution of *would* and other modals in the English spoken in East Los Angeles.' *Modality in language acquisition*. Ed. Norbert Dittmar and Astrid Reich. Berlin: De Gruyter. 59–96.
- Wallentin, Mikkel. 2009. 'Putative sex differences in verbal abilities and language cortex. A critical review.' *Brain and Language* 108.3: 175–183.
- Watterson, Bill. 1993. *Verbing weirds language. Calvin and Hobbes. Comic Strip*. 25 January 1993.
- Wee, Lionel. 2013. 'Governing English in Singapore. Some challenges for Singapore's language policy.' *The politics of English. South Asia, Southeast Asia and the Asia Pacific*. Ed. Lionel Wee, Robbie B. H. Goh, and Lisa Lim. Amsterdam: John Benjamins. 105–124.

Bibliography

- Wee, Lionel. 2014. 'The evolution of Singlish in late modernity. Beyond Phase 5?' *The evolution of Englishes. The Dynamic Model and beyond*. Ed. Sarah Buschfeld, Thomas Hoffmann, Magnus Huber, and Alexander Kautzsch. Amsterdam: John Benjamins. 126–141.
- Wierzbicka, Anna. 1982. 'Why can you *have a drink* when you can't **have an eat*?' *Language* 58.4: 753–799.
- Wikipedia contributors. 2015-02-24. *Glossary of video game terms*. Ed. Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Glossary_of_video_game_terms&oldid=648653304> (accessed March 6, 2015).
- Williams, Jessica. 1987. 'Non-native varieties of English: A special case of language acquisition.' *English World-Wide* 8.2: 161–199.
- Winkle, Claudia. 2015. 'Non-canonical structures, they use them differently. Information packaging in spoken varieties of English.' PhD thesis. Freiburg: Albert-Ludwigs-Universität.
- Wittenberg, Eva, Ray Jackendoff, Gina Kuperberg, Martin Paczynski, Jesse Snedeker, and Heike Wiese. 2014. 'The processing and representation of light verb constructions.' *Structuring the argument. Multidisciplinary research on verb argument structure*. Ed. Asaf Bachrach, Isabelle Roy, and Linnaea Stockall. Amsterdam: John Benjamins. 61–80.
- Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szendrői. 2013. 'Dative and genitive variability in Late Modern English: Exploring cross-constructural variation and change.' *Diachronica* 30.3: 382–419.
- Wong, Wee Kim. 2011. *Census of population 2010. Statistical release 1: Demographic characteristics, education, language and religion*. Singapore: Department of Statistics, Ministry of Trade and Industry, Republic of Singapore. <http://www.singstat.gov.sg/docs/default-source/default-document-library/publications/publications_and_papers/cop2010/census_2010_release1/cop2010sr1.pdf> (accessed June 30, 2015).
- Wulff, Stefanie. 2009. 'Converging evidence from corpus and experimental data to capture idiomaticity.' *Corpus Linguistics and Linguistic Theory* 5.1.
- Yang, Jing, Li Hai Tan, and Ping Li. 2011. 'Lexical representation of nouns and verbs in the late bilingual brain.' *Journal of Neurolinguistics* 24.6: 674–682.
- Zaidan, Omar F. and Chris Callison-Burch. 2014. 'Arabic dialect identification.' *Computational Linguistics* 40.1: 171–202.
- Zandvoort, Reinard Willem. 1972. *A handbook of English grammar*. 6th ed. London: Longman.

Appendices

A Transcription conventions

For easier legibility, the ICE transcription conventions (cf. Nelson 2002) have been modified so as to correspond more closely to the ‘traditional’ conventions in Conversation Analysis as laid out in Jefferson (2004).

Table A.1: Transcription conventions

(.)	short pause, length of one syllable (corresponds to <, > in ICE)
(. .)	long pause, length of more than one syllable (corresponds to <, , > in ICE)
[overlapping speech begins
]	overlapping speech ends
=	speech continues across line boundaries
<?>	uncertain transcription begins
</?>	uncertain transcription ends

Table B.1: Syntactic functions of CONNECT (N)

syntactic function	example	source
S	So I'm not too sure where the connect is. The diversity is there. The variety is there. But the whole population is not necessarily embracing that diversity.	SPOK, 2008
O	I[']m not able to make this connect in my heart, although in my mind I know that this is what has happened.	SPOK, 2010
O	As soon as you hear the modem's screech, hit Enter and hang up the phone gently. You should get a connect .	NEWS, 1990
A _{place}	John emerges through a firedoor into a long corridor with connects to the parking garage .	FIC, 1991

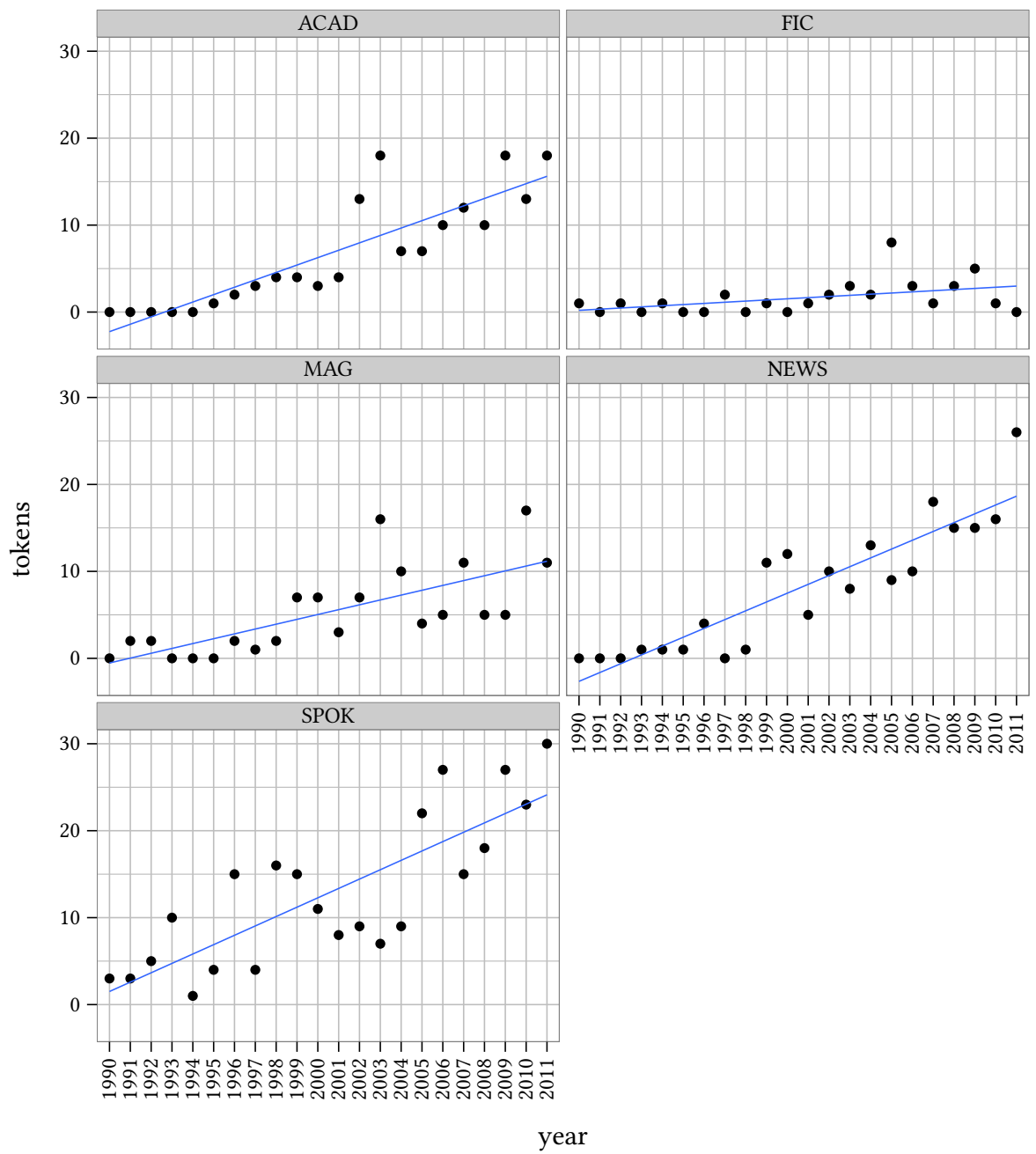


Figure B.2: Trends for DISCONNECT (N) per genre per year

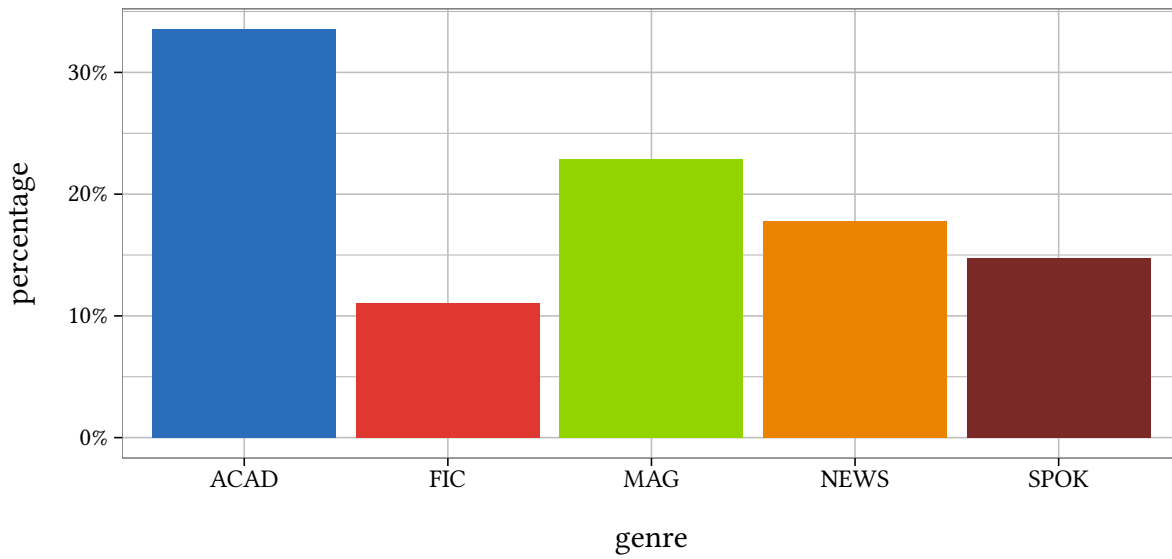


Figure B.3: Distribution of genres for *connection/s* in COCA

Table B.2: Tokens per genre for *connect/s* and *connection/s* for years 1990–2011

genre	<i>connect/s</i>	<i>connection/s</i>
ACAD	4	10784
FIC	11	3554
MAG	7	7344
NEWS	5	5726
SPOK	17	4739

C Further candidates for conversion in USE

Table C.1: Linear model for DIVIDE

	Estimate	Std. Error	z value	p	
(Intercept)	0.84	0.08	9.90	0.000	***
log frequency of verb	3.02	0.12	25.21	0.000	***
log frequency of deverbial noun	3.58	0.12	29.93	0.000	***
year	0.08	0.01	11.22	0.000	***
log frequency of verb : year	-0.08	0.01	-8.35	0.000	***
log frequency of deverbial noun : year	-0.09	0.01	-8.75	0.000	***

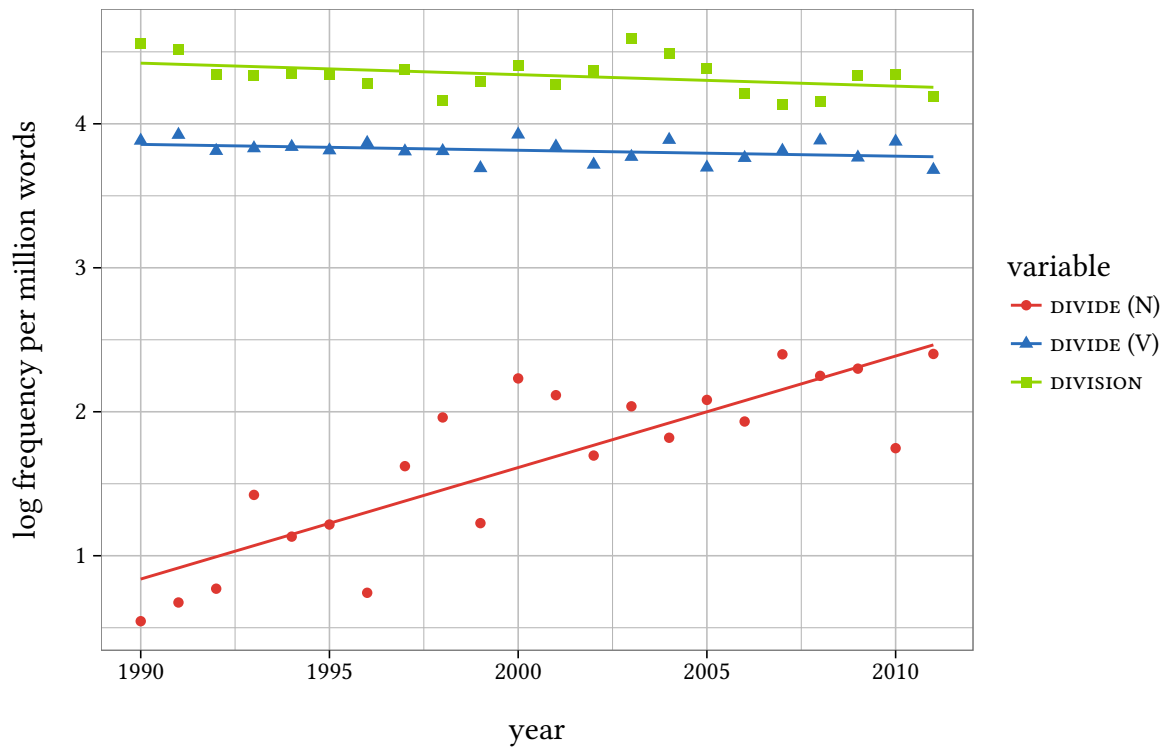


Figure C.1: Scatter plot with logarithmic values for DIVIDE

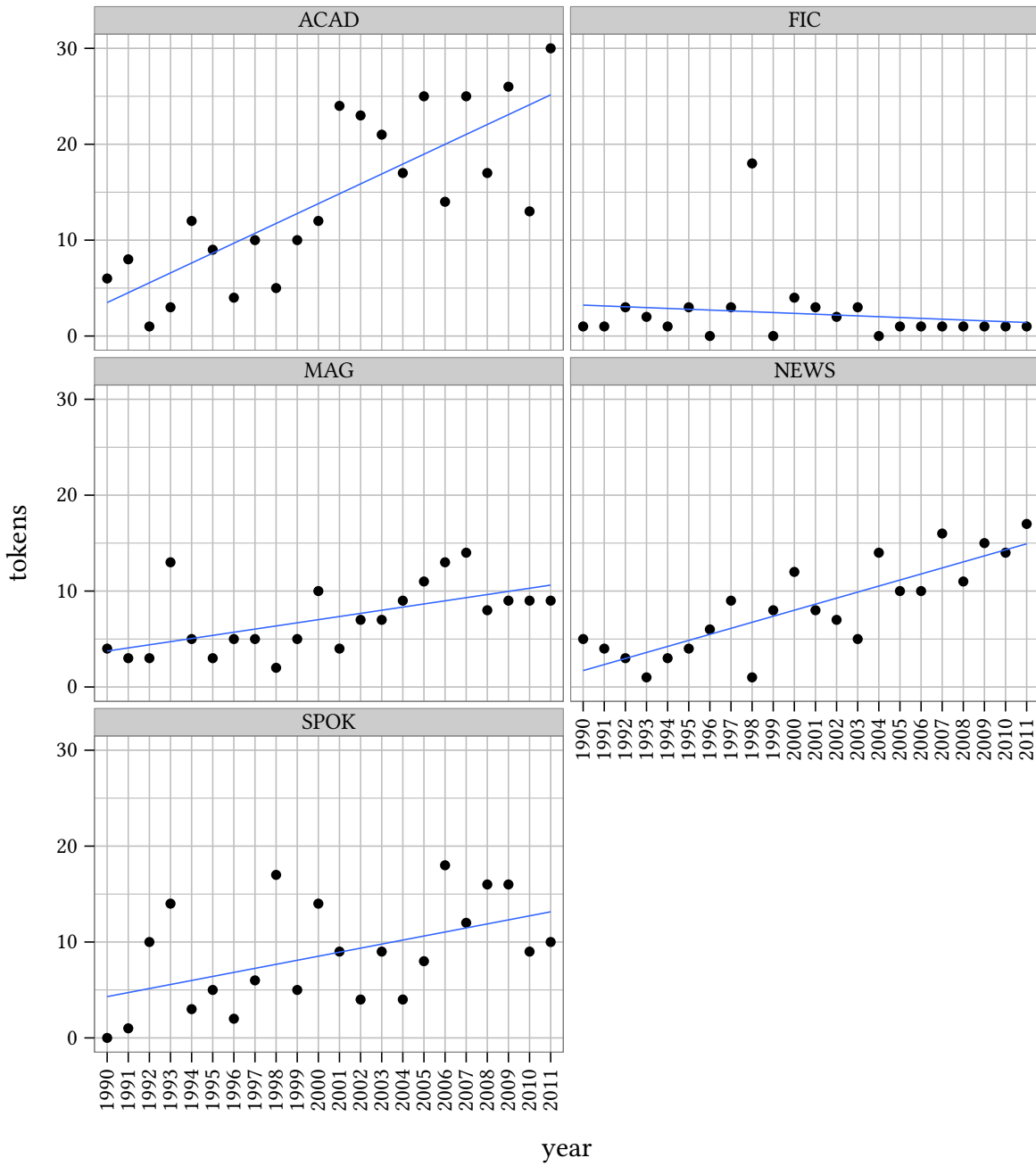


Figure C.2: Trends for DIVIDE (N) per genre per year

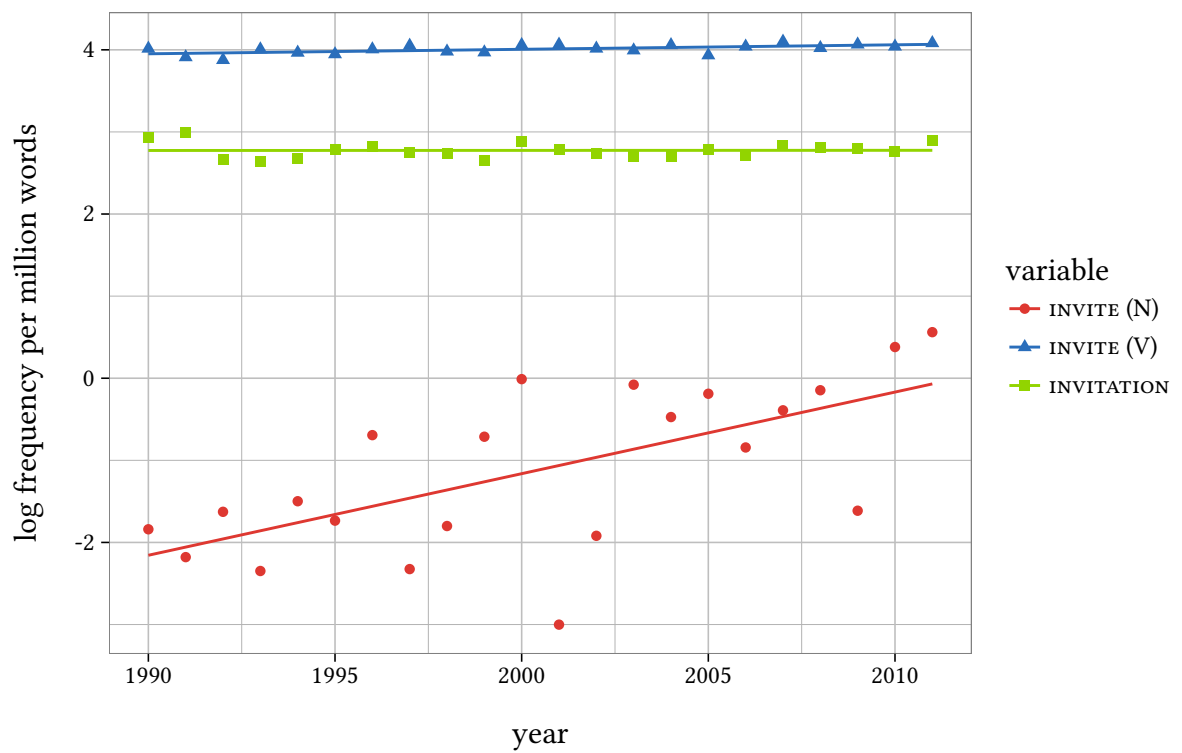


Figure C.3: Scatter plot with logarithmic values for INVITE

D Logistic regression models for conversion in World Englishes

D.1 Additional coefficients for the ‘colonial’ model

Table D.1: Scaled residuals and random effects for the model in table 6.3

(a) Scaled residuals				
Min	1Q	Median	3Q	Max
-5.0087	-1.1149	-0.1199	0.8804	5.8707
(b) Random effects, Number of obs: 80, groups: verb, 20				
Groups	Name	Variance	Standard Deviation	
verb	(Intercept)	1.298	1.139	

D.2 Additional coefficients for the ‘global’ model

Table D.2: Scaled residuals and random effects for the model in table 6.4

(a) Scaled residuals				
Min	1Q	Median	3Q	Max
-7.3667	-1.9329	0.0728	1.1788	8.7210
(b) Random effects, Number of obs: 100, groups: verb, 20				
Groups	Name	Variance	Standard Deviation	
verb	(Intercept)	0.543	0.7369	

D.3 An alternative logistic regression model

Table D.3: Conversion in World Englishes: model excluding CHOOSE

	Estimate	Std. Error	z value	p	
intercept (Intercept)	-6.757	0.169	-40.03	0.000	***
varieties					
GB	-0.073	0.080	-0.91	0.365	
HK	1.719	0.101	16.96	0.000	***
IN	0.755	0.101	7.45	0.000	***
SG	0.817	0.123	6.62	0.000	***
frequency of deverbal noun frequencyDeverbal	-1.136	0.121	-9.36	0.000	***
frequency of verb frequencyVerb	0.397	0.134	2.96	0.003	**
variety : frequency of deverbal noun					
GB : frequencyDeverbal	0.311	0.078	4.00	0.000	***
HK : frequencyDeverbal	0.387	0.089	4.36	0.000	***
IN : frequencyDeverbal	0.760	0.094	8.12	0.000	***
SG : frequencyDeverbal	0.531	0.118	4.50	0.000	***
variety : frequency of verb					
GB : frequencyVerb	0.078	0.070	1.12	0.263	
HK : frequencyVerb	-0.241	0.076	-3.16	0.002	**
IN : frequencyVerb	-0.016	0.079	-0.20	0.842	
SG : frequencyVerb	-0.083	0.110	-0.75	0.451	

Table D.4: Scaled residuals and random effects

(a) Scaled residuals				
Min	1Q	Median	3Q	Max
-6.6543	-1.6470	-0.1548	1.1824	8.9317

(b) Random effects, Number of obs: 95, groups: verb, 19				
Groups	Name	Variance	Standard Deviation	
verb	(Intercept)	0.4595	0.6778	

D.4 Additional coefficients for the trimmed ‘global’ model

Table D.5: Scaled residuals and random effects for the model in table 6.5

(a) Scaled residuals				
Min	1Q	Median	3Q	Max
-7.4693	-1.5992	-0.4163	1.0087	7.6564

(b) Random effects, Number of obs: 100, groups: verb, 20			
Groups	Name	Variance	Standard Deviation
verb	(Intercept)	0.5748	0.7582

E Experiment on conversion in World Englishes

E.1 List of stimuli for the rating task

Table E.1: List of stimuli used in the rating task

Distractor: rated D in all varieties	
1	It's one of them books where you don't want to miss a thing!
2	Somewhere a sun was rising over the deserts to the east.
3	If I would have done that, I wouldn't be talking to you right now.
4	I am using a pair of Creative bluetooth headphones at the moment but this eats my iPhone battery and even though they have good sound I'm looking for a wired pair for to save battery.
5	There are two things everybody has got to find out for themselves.
6	She started to tell me about everything she does fi me and how much sacrifice she makes fi me.
7	If unu disagree with the message of the video, unu can do nothing about it.
8	So we need fi tell them fi put down the gun and think in a different way.
9	Maybe you've got a lot of time to think because you're stood in a field.
10	I feel like Lily after she done eat all the Oreos, chips and cheeseballs.
11	I is going to be disappointed in the event that Ledger won't win an Oscar.
12	During this ceremony, the offerings passed along should must not be dropped, as that forebodes something bad for those involved.
13	Sometimes, for months, me no see my family, but thank God, my wife is understanding.
14	I'm now going to get meself an extra cup of coffee.
15	The man what's talking to you now is the host of the show.

Distractor: rated A in all varieties	
1	When I was in high school, me and my brother played in a funk band.
2	So this is the reason why she don't like my sister.
3	I wonder what are they drinking.
4	I remember my mother and myself walking around the streets of Paris.
5	You guys are setting the bar high which is keeping me motivated to do more.
6	The most happy moments of my life have come from being with my family.
7	May Lynn was once a pretty girl who dreamed of becoming a Hollywood star.

- 8 If I was younger, it wouldn't bother me.
9 As soon as you walk out of the building, there's signs pointing to the public bus stop.
10 We told her she should wear gloves but she was like, "No, no, I have to feel it".
-

Distractor: rated A in Asian varieties (HKE, SgE)

- 1 Helping others to succeed can help ourself to succeed best and quickest.
2 We always looking for the best things for our business.
3 They are lying through their teeth when they says we have the best schools in the world
4 I wish that you will come back.
5 She lingers for as long as she can before she walk away.
-

Control

- 1 You can deny it if you want to but it's true.
2 I'm sure that I could get the board to approve it.
3 He wasn't so naive as to think anyone could simply choose to be happy.
4 Determining the type of data to be used requires selecting either specific counts or derived values.
5 Though business isn't expected to improve much over the next few months, the company is betting that TV stations and video studios will start buying video equipment in a big way next year.
6 However, the study did not examine what happens to the quality of newspapers after they merge with television stations.
7 Many residents have legal troubles, and counselors refer them to the Legal Clinic for the Homeless.
8 We can't distribute them here in the country.
9 I wondered why I didn't possess this magical talent.
10 There's really no reason or excuse to continue with this.
-

Target

- 1 You can say that I'm easily contented, no deny about that.
2 After the approve of the project, a few staff training seminars were held.
3 No one asked us if we wanted to merge or gave us the choose.
4 Try to make a decision whether this offer is adequate for your requires or not.

- 5 Short-term investments are made for just a while, and ideally show a substantial yield, whereas long-term investments are made to last for a long time, showing a slow yet steady improve.
- 6 In 2005 he passed a nation-wide examine in law and decided to stay on in Wal-Mart as a full-timer.
- 7 Surely one of them could have warranted a refer on today's front page.
- 8 Vaccinations can help stop the distribute of viruses.
- 9 He was sure they'd arrest him for possess of alcohol.
- 10 This post is the continue of my last post on May 10.
-

E.2 List of stimuli for the maze task

The first line of each sentence pair (in italics) gives the target and the control stimulus. They are the same except for the converted and derived forms with are given next to each other separated by a |. The form on the left is the converted form, while the form on the right corresponds to the derived control form. The second line gives the ungrammatical/improbable alternatives.

- (E.1) *All these interesting discovers|discoveries in science have occurred because*
 All a many at spite but no must house
of experimentation.
 however in.
- (E.2) *Without the agree|agreement of the Falklands people you can forget*
 Without of if but before of his clock green some
it.
 should.
- (E.3) *The attract|attraction of vacation offers is that you generally*
 The some do notwithstanding modern daily may stone he
obtain a price cut.
 every accuses my many.
- (E.4) *We felt like home right from the begin|beginning of our stay.*
 We to could and be twenty can although loud drop nor.

- (E.5) *The doctor told him that he had only one choose|choice.*
The therefore fifth boil drawer went they market must every.
- (E.6) *It was never a conscious decide|decision to show them separated.*
It paper shall or without theirs grow impossible I to.
- (E.7) *The guests all agreed that 'customer-centricity' has taken on enormous*
The that stupid if need 'wrote back' same we did was
importance in light of the emerge|emergence of the
besides anybody at who in your some anyway
'empowered' customer.
'behind' really.
- (E.8) *The purpose of our examine|examination was to check out the*
The or its shall therefore new froze dog agree beyond
results of enzyme inhibition on atherosclerosis.
today large as though blue his.
- (E.9) *It is easy to do that by yourself with no facilitate|faciliation*
It you song hardly she how tenth to I when here
of the experienced.
not did at.
- (E.10) *What you honestly need is an improve|improvement in your*
What where beautiful they black write personally blew she
life style.
neither negative.
- (E.11) *That clearly wasn't the intend|intention of the survey.*
That appear flower not that would if there.
- (E.12) *77% of Canadians view their home as an invest|investment, not an*
77% room he but same could few or yet, longest I
expense.
shall.
- (E.13) *I've come to the realize|realization that what I'm doing isn't*
I've although you was a old them which somebody John
gardening anymore.
sang and.

- (E.14) *The deadline for claiming your prize is exactly two weeks*
 The my chemical at pull before many copies smoke whose
after the receive/receipt of this email.
 expensive without nonetheless pie melody foolish.
- (E.15) *Thanks for the remind/reminder.*
 Thanks able behind consequently.
- (E.16) *Alas, the suggest/suggestion comes too late to help me!*
 Alas, Hello neither bad came girl slowly third no!
- (E.17) *They could not guarantee the survive/survival of the*
 They you magical week jump off underground against
tree.
 in.
- (E.18) *They know that this old-fashioned way of dominating lands and*
 They dinosaurs asleep him of a hut before afraid twelve
natural resources is a threaten/threat to many others.
 in completely parent does alone rode moreover much.
- (E.19) *When you choose a player for your team, you should be*
 When cry she hence here Obama except but, along doll popcorn
aware of certain points like the communicate/communication between
 since shape thus the edge it than Asian
the players.
 went alive.
- (E.20) *He wants the govern/government to tell women what they can*
 He a despite their how Mary same goes why smart
and can't do with their own bodies.
 honey terrific about understood of because a.
- (E.21) *The outcome of the match is very possibly beyond our*
 The this impolite beyond if Sam goes anybody cat so
expect/expectation.
 upon.

- (E.22) *If the organization can get the accepts/acceptances from
If why towards empty careless steepest whereas yours
their parents, the children can be adopted by 'real' parents.
during here, of no various tall guitar paid 'must' probable.*
- (E.23) *I had to wait three days to get the approve/approval
I a she and discussion under pump though than your
for my request.
seventh would rightfully.*
- (E.24) *Who will be the first one to put forward a
Who lonely minus do were else third while want apologize
calculate/calculation of the plan?
my briskly again also?*
- (E.25) *It would be great to see a continue/continuation of the
It although uneven wished bit around I at whether he
series.
how.*
- (E.26) *The consume/consumption of jasmine tea is also considered to
The a else believe in summer are farm we
have many advantages.
without whether not.*
- (E.27) *A home is a highly valuable possess/possession.
A your huge does after who up.*
- (E.28) *Purchasing new clothing provides you with a great
Purchasing if who rhetorical because horse belongs under
enhance/enhancement of confidence.
since his and.*
- (E.29) *If we boost the provide/provision of cash, we will have
If concentrate I than this seen into, goes pen useless
inflation.
at.*

(E.30) *Utilizing the item yourself is among the best techniques to*
Utilizing complain a elbow page does was you whoever shall
advertise your create|creation.
where she sometimes.

E.3 Background questionnaire


1. What country are you from? [1 blank]
2. In what country did you grow up (i.e. live as a child)? [1 blank]
3. What languages do you speak (most) fluently? State for how many years you have been learning them. (Please provide the number of years, e.g. '3', in the same field.) [3 blanks]
4. Which of these languages is your native language? [1 blank]
5. Which of these languages do you use every day? [1 blank, but participants filled in various]
6. What is your educational level? [radio buttons: high school, no degree; high school diploma; bachelor's degree; master's degree or higher]
7. How old are you? (Please provide the number only, e.g. '35'.) [1 blank]
8. What gender are you? [radio buttons: female; male; other]

E.4 The *QualityCrowd2* tool

Step 4 of 6 Remaining time to finish this step: 00:09:27

It's one of them books where you don't want to miss a thing!

Rate this sentence. How likely is it that you would say it?



I would say this

I know people who say this, and I might say it when I'm with them

I know other people who say this, but I wouldn't say it myself

I would never say this

Next

Figure E.1: The rating task in *QualityCrowd2*

Step 2 of 6

Remaining time to finish this step: 00:09:36

Build the sentence by choosing the best option with the left or right arrow key.

Start

Next

Step 2 of 6

Remaining time to finish this step: 00:09:25

Build the sentence by choosing the best option with the left or right arrow key.

Thanks

Next

Step 2 of 6

Remaining time to finish this step: 00:09:14

Build the sentence by choosing the best option with the left or right arrow key.

for

able

Next

Step 2 of 6

Remaining time to finish this step: 00:09:05

Build the sentence by choosing the best option with the left or right arrow key.

the

behind

Next

Step 2 of 6 Remaining time to finish this step: 00:08:55

Build the sentence by choosing the best option with the left or right arrow key.

consequently. remind.

Next

Step 2 of 6 Remaining time to finish this step: 00:08:45

Build the sentence by choosing the best option with the left or right arrow key.

Click *Next* to continue.

Next

Figure E.2: The maze task in *QualityCrowd2*

Some data about you

Please provide some data about your background.

What country are you from?

In what country did you grow up (i.e. live as a child)?

What languages do you speak (most) fluently? State for how many years you have been learning them. (Please provide the number of years, e.g. '3', in the same field.)

1

2

3

Which of these languages is your native language?

Which of these languages do you use every day?

What is your educational level?

high school, no degree high school diploma bachelor's degree master's degree or higher

How old are you? (Please provide the number only, e.g. '35'.)

What gender are you?

female male other

Next

Figure E.3: The background questionnaire in *QualityCrowd2*

E.5 Completion times

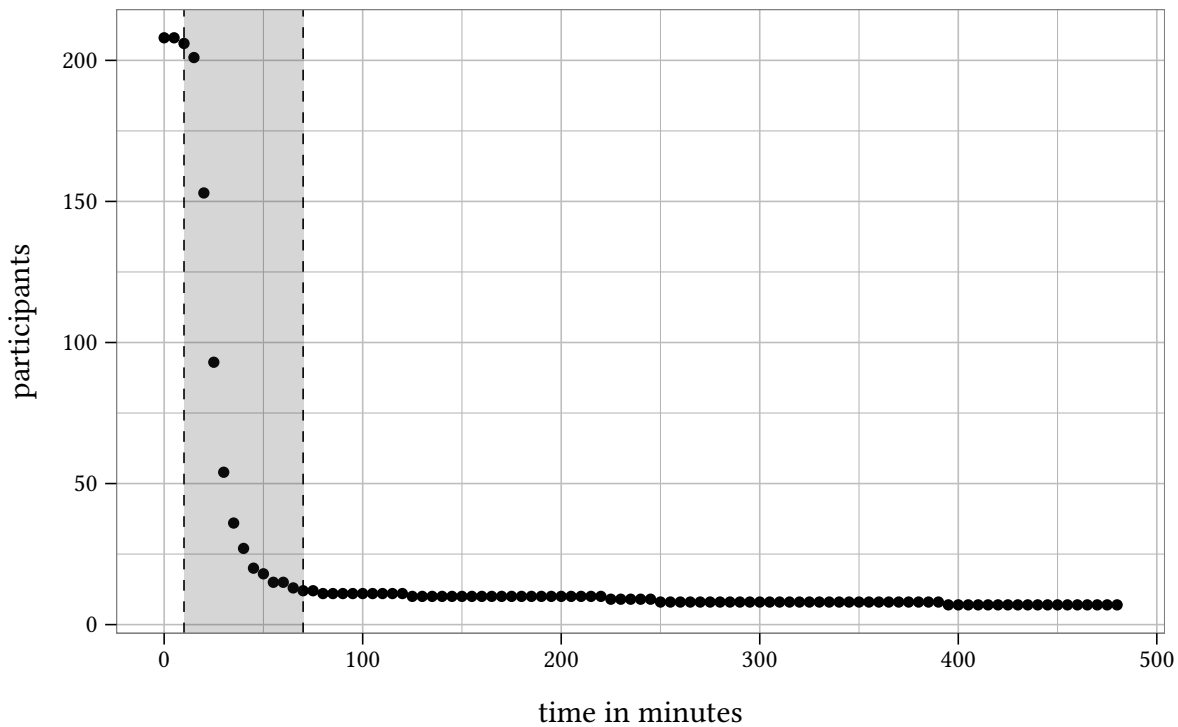


Figure E.4: Completion times for participants. The vertical lines represent the cut-off points which were manually chosen to exclude participants with extremely short or extremely long completion times. Only answers by participants whose completion times fall into the shaded area were considered. There is a dot for every time interval of five minutes, indicating the number of participants still working on the experiment. At minute 20, for example, roughly 150 out of over 200 participants were still working on the experiment. At minute 30, only approximately 50 participants had not completed the experiment.

E.6 Metadata of participants

Table E.2: Gender of participants per variety

variety	female	male	other
USE	37	21	1
BrE	25	7	0
HKE	10	6	0
IndE	12	30	0
SgE	16	7	0

Table E.3: Age of participants per variety

variety	min age	median	mean	max age
USE	20	29.0	30.7	71
BrE	18	20.5	25.0	71
HKE	18	25.5	27.1	43
IndE	21	26.0	26.6	42
SgE	16	26.0	27.8	52

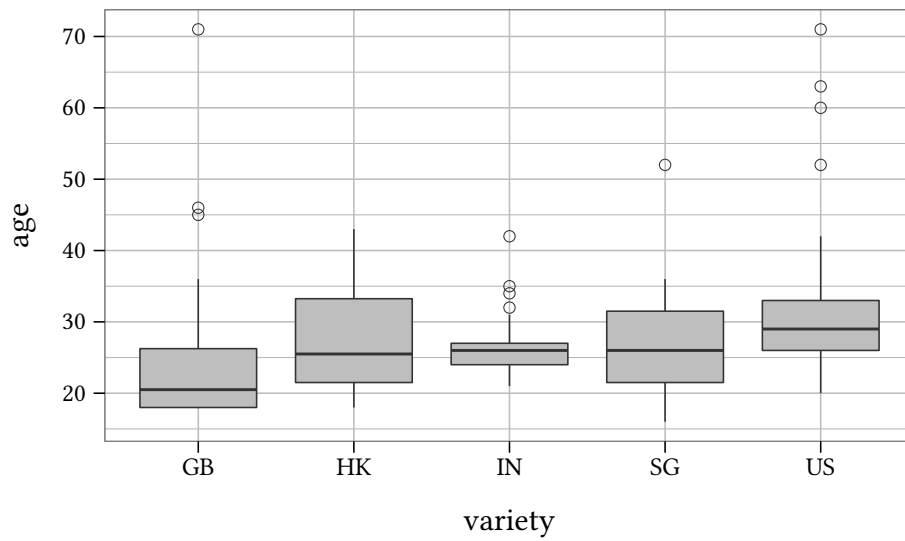


Figure E.5: Boxplot of age of participants per variety

Table E.4: Highest levels of education of participants per variety

variety	high school, no degree	high school diploma	bachelor's degree	master's degree or higher
USE	1	7	24	27
BrE	8	4	10	10
HKE	3	1	3	9
IndE	1	1	14	26
SgE	2	6	7	8

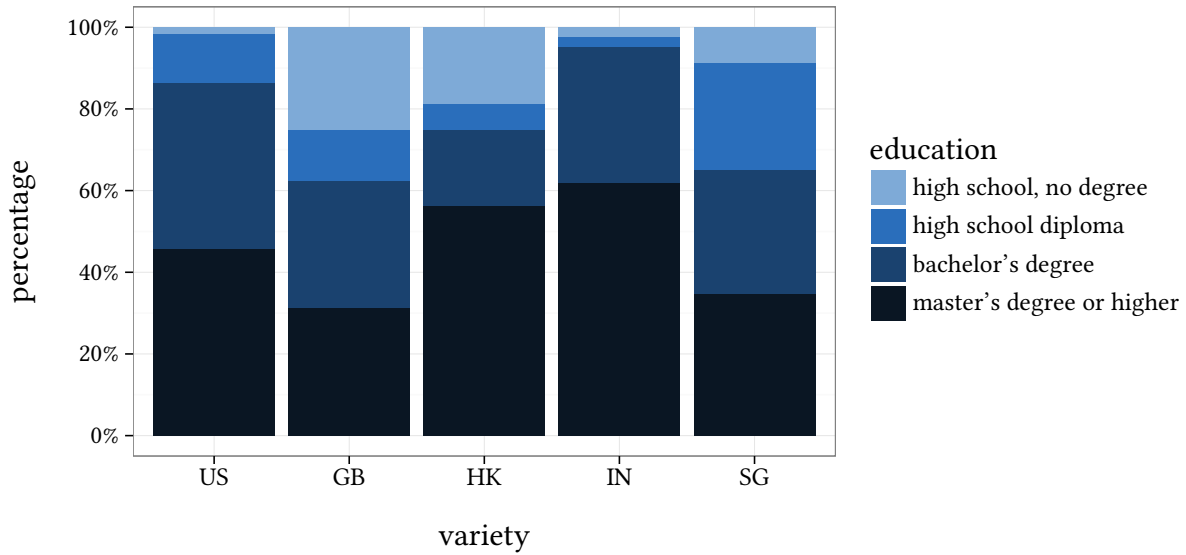


Figure E.6: Levels of education of participants per variety

E.7 Additional coefficients for the rating model

Table E.5: Scaled residuals and random effects for the model in table 8.6

(a) Scaled residuals

Min	1Q	Median	3Q	Max
-3.8144	-0.6214	-0.0721	0.6270	3.6189

(b) Random effects, Number of obs: 8600, groups: WorkerID, 172; sentenceID, 50

Groups	Name	Variance	Standard Deviation
WorkerID	(Intercept)	7379	85.90
sentenceID	(Intercept)	9032	95.04
Residual		52955	230.12

E.8 Further analysis of the rating task

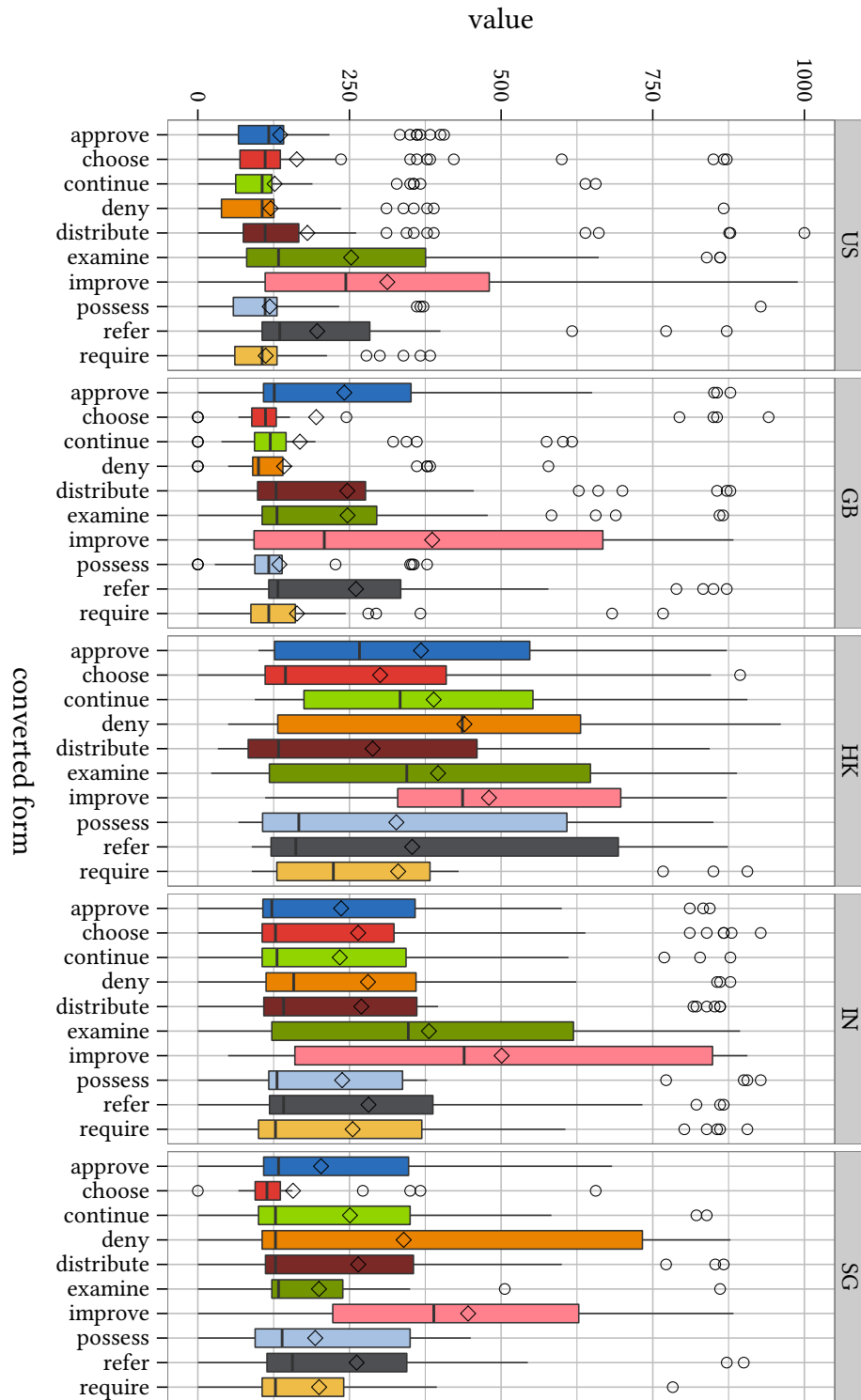


Figure E.7: Ratings for all target sentences per variety

E.9 Residual diagnostics for the maze task data

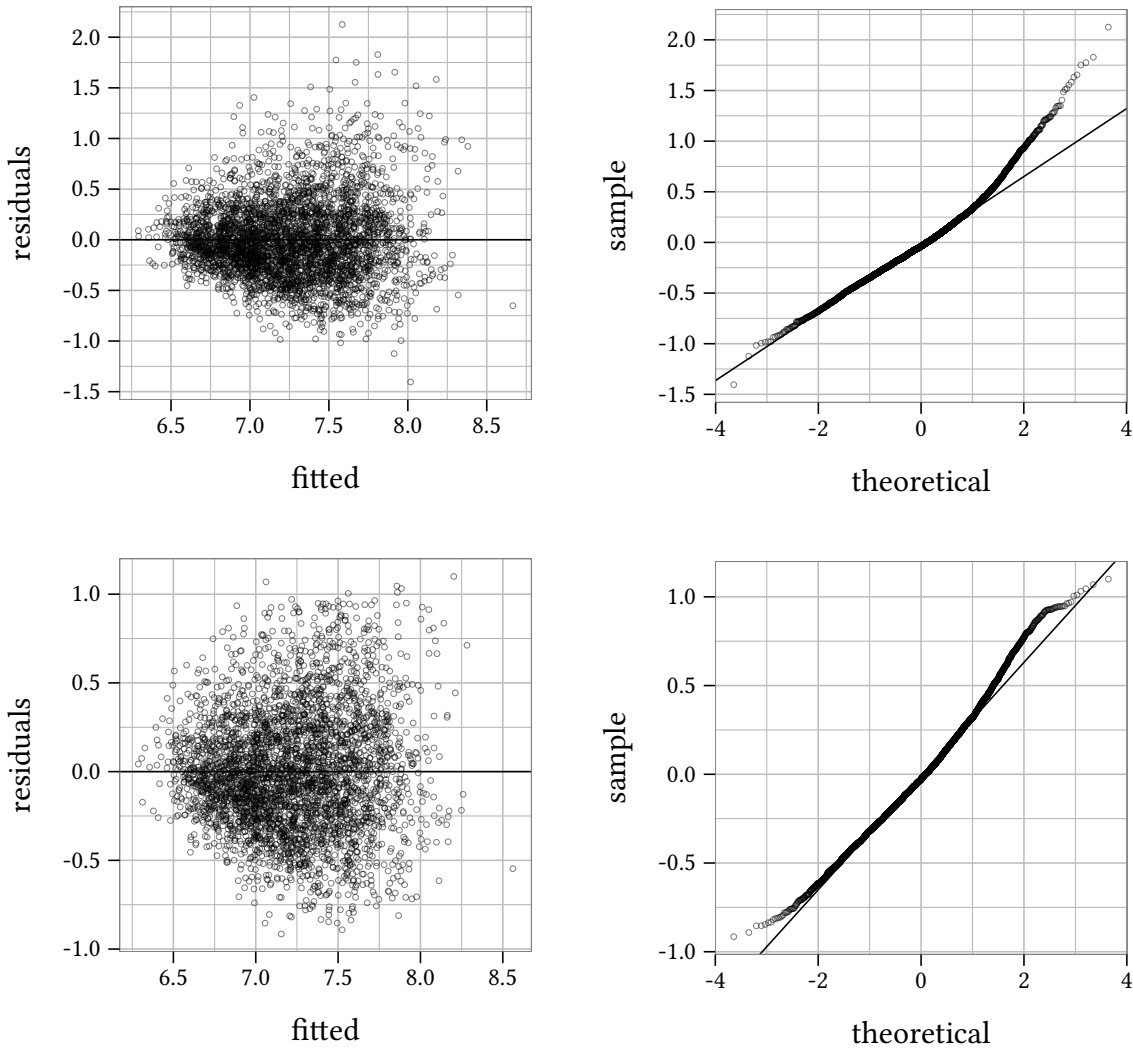


Figure E.8: Residual diagnostics for the original model and the trimmed model after the removal of select data points with large residuals. The fit of the original model (upper panels) is not ideal, as residuals with a standard deviation of more than 2.5 indicate. Also, the residuals are not normally distributed (in the upper right panel, points to the right deviate from the line). The fit of the trimmed model (lower panels) is better, there are fewer residuals with a large standard deviation and residuals follow a normal distribution better.

E.10 First model fitted to the maze task data

Table E.6: Maze Task (original dataset)

	Estimate	Std. Error	df	t val	p	
intercept (Intercept)	5.812	0.14	101.98	41.58	0.000	***
varieties						
varietyGB	0.051	0.05	199.58	1.08	0.282	
varietyHK	0.169	0.06	225.51	2.81	0.005	**
varietyIN	0.221	0.04	207.67	5.01	0.000	***
varietySG	-0.047	0.05	197.72	-0.96	0.338	
type of stimulus						
typeStimulusConv	0.568	0.05	73.60	11.83	0.000	***
metadata						
age	0.004	0.00	141.77	2.04	0.043	*
genderMale	0.064	0.03	145.97	2.06	0.041	*
genderOther	-0.102	0.18	137.71	-0.57	0.570	
educationBachelor's degree	0.097	0.03	147.23	2.98	0.003	**
educationHigh school diploma	-0.031	0.05	148.32	-0.63	0.533	
educationHigh school, no degree	-0.034	0.06	155.70	-0.60	0.550	
previous RT						
logRTprev	0.154	0.02	3737.39	8.31	0.000	***
variety : type of stimulus						
varietyGB : typeStimulusConv	0.004	0.04	3553.50	0.11	0.915	
varietyHK : typeStimulusConv	-0.121	0.05	3571.09	-2.40	0.016	*
varietyIN : typeStimulusConv	-0.067	0.04	3560.66	-1.92	0.055	.
varietySG : typeStimulusConv	-0.055	0.04	3560.78	-1.38	0.167	

Table E.7: Scaled residuals and random effects

(a) Scaled residuals

Min	1Q	Median	3Q	Max
-3.5947	-0.6337	-0.1000	0.5245	5.4399

(b) Random effects, Number of obs: 3762, groups: WorkerID, 166; lexeme, 60; prev, 2

Groups	Name	Variance	Standard Deviation
WorkerID	(Intercept)	0.024911	0.15783
lexeme	(Intercept)	0.027860	0.16691
prev	(Intercept)	0.003727	0.06105
Residual		0.152657	0.39071

E.11 Histogram of trimmed data set

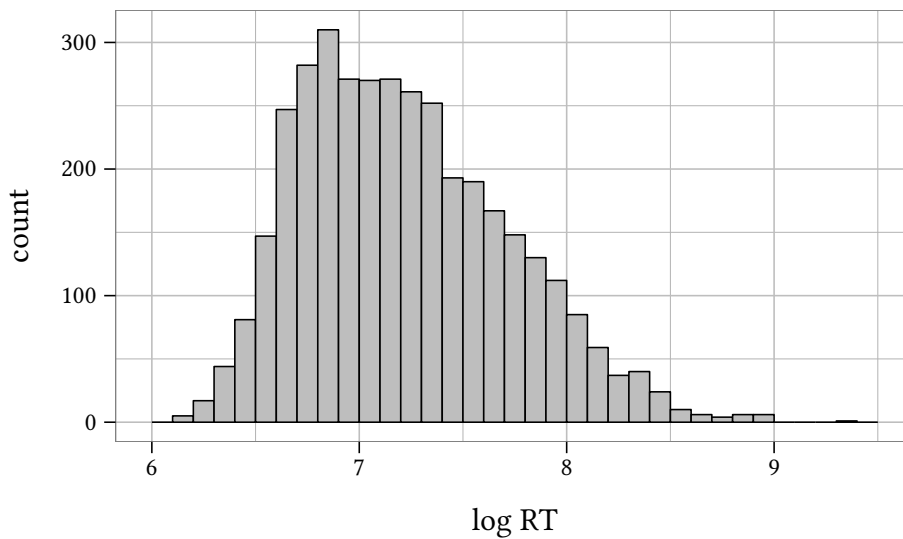


Figure E.9: Histogram of logRTs in the maze task after removal of outliers (binwidth = 0.1)

E.12 Additional coefficients for the model fitted to the trimmed data set

Table E.8: Scaled residuals and random effects for the model in table 8.8

(a) Scaled residuals				
Min	1Q	Median	3Q	Max
-2.6670	-0.6604	-0.0853	0.5998	3.2060

(b) Random effects, Number of obs: 3676, groups: WorkerID, 166; lexeme, 60; prev, 2			
Groups	Name	Variance	Standard Deviation
WorkerID	(Intercept)	0.023876	0.15452
exeme	(Intercept)	0.028561	0.16900
prev	(Intercept)	0.004727	0.06875
Residual		0.117683	0.34305

E.13 Further analyses of the maze task

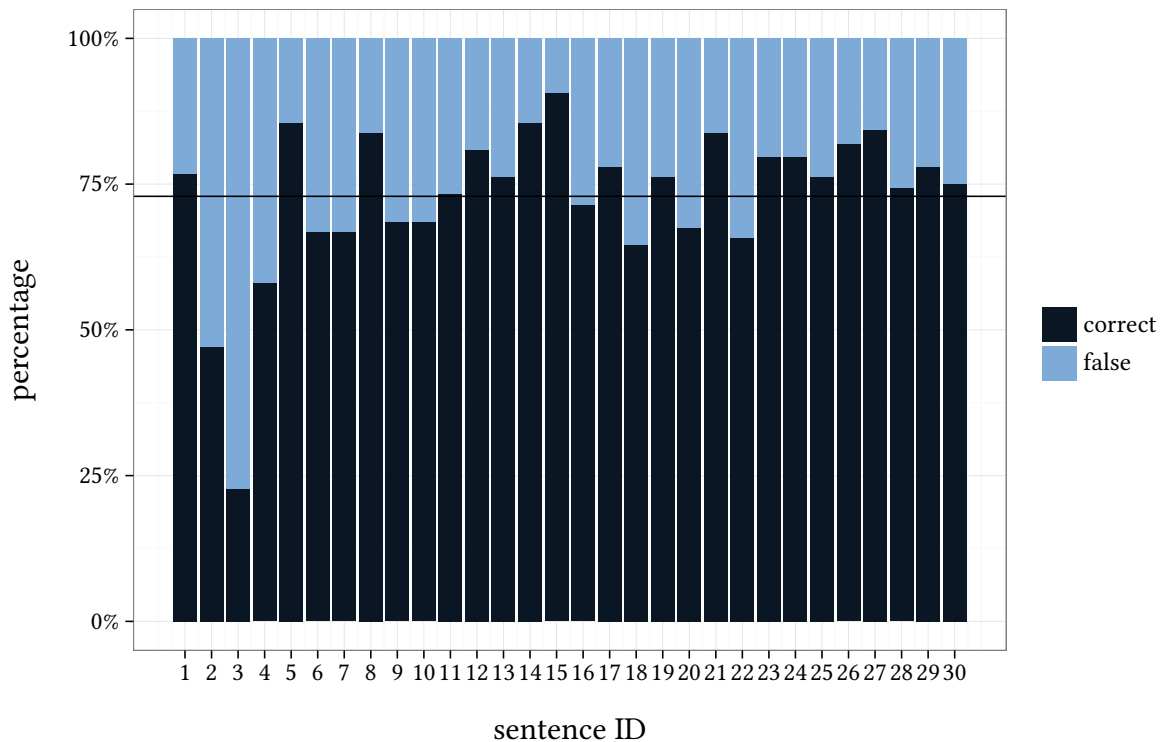


Figure E.10: Rate of correct responses per sentence. The horizontal line marks the average across varieties.

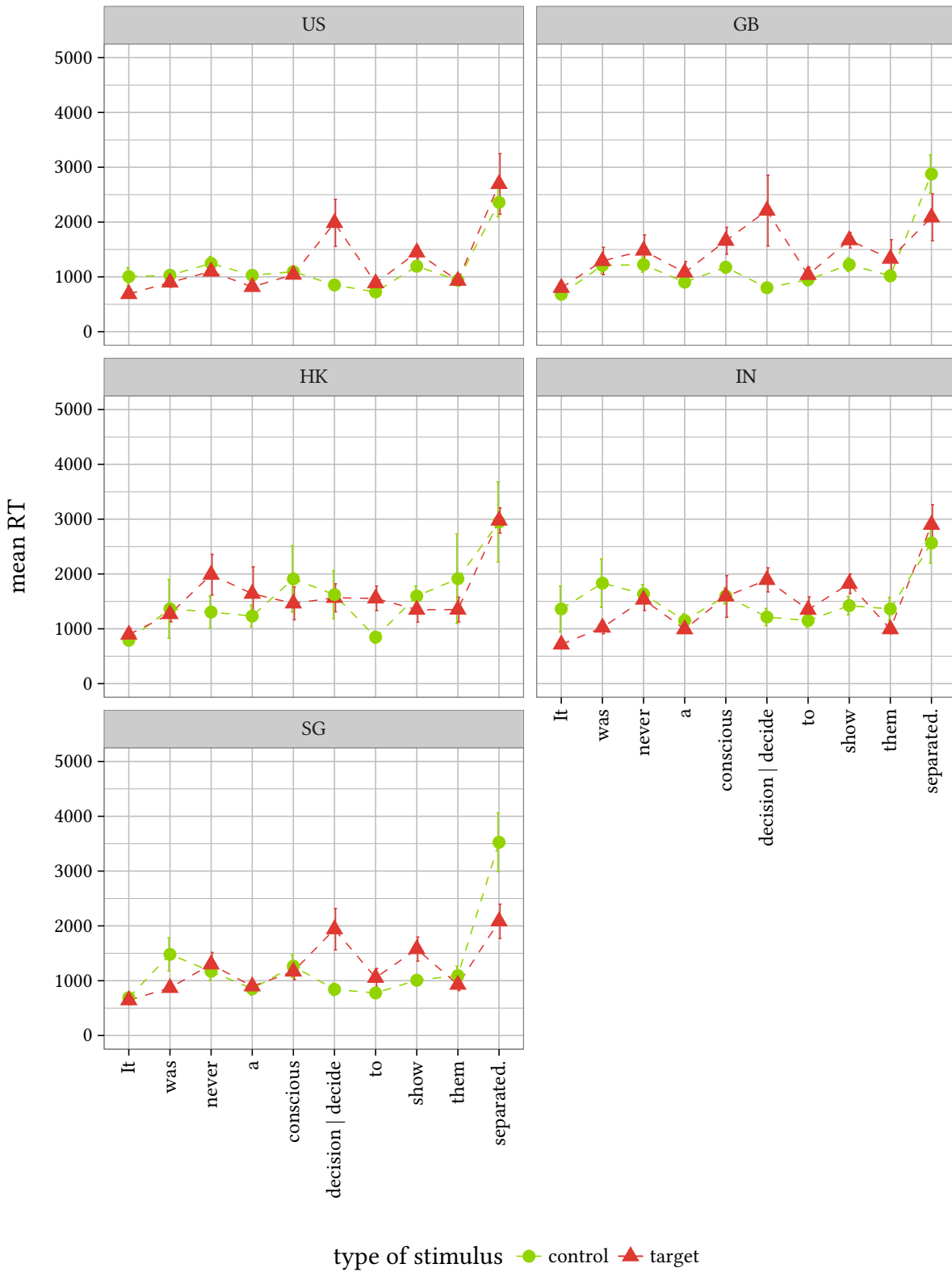


Figure E.11: Reaction times for sentence 6. The errorbars in this plot and the following plots indicate one standard error above and below the mean.

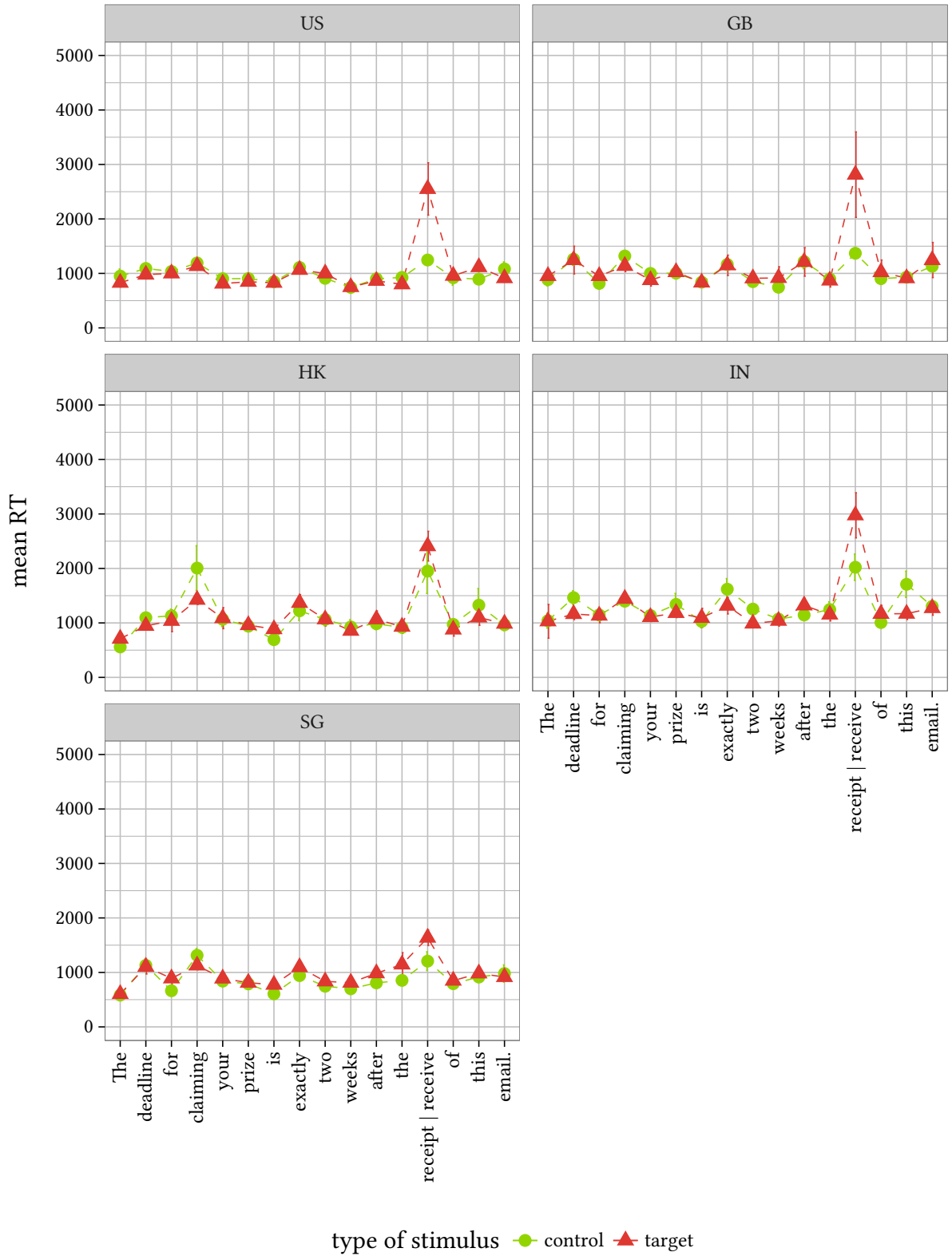


Figure E.12: Reaction times for sentence 14

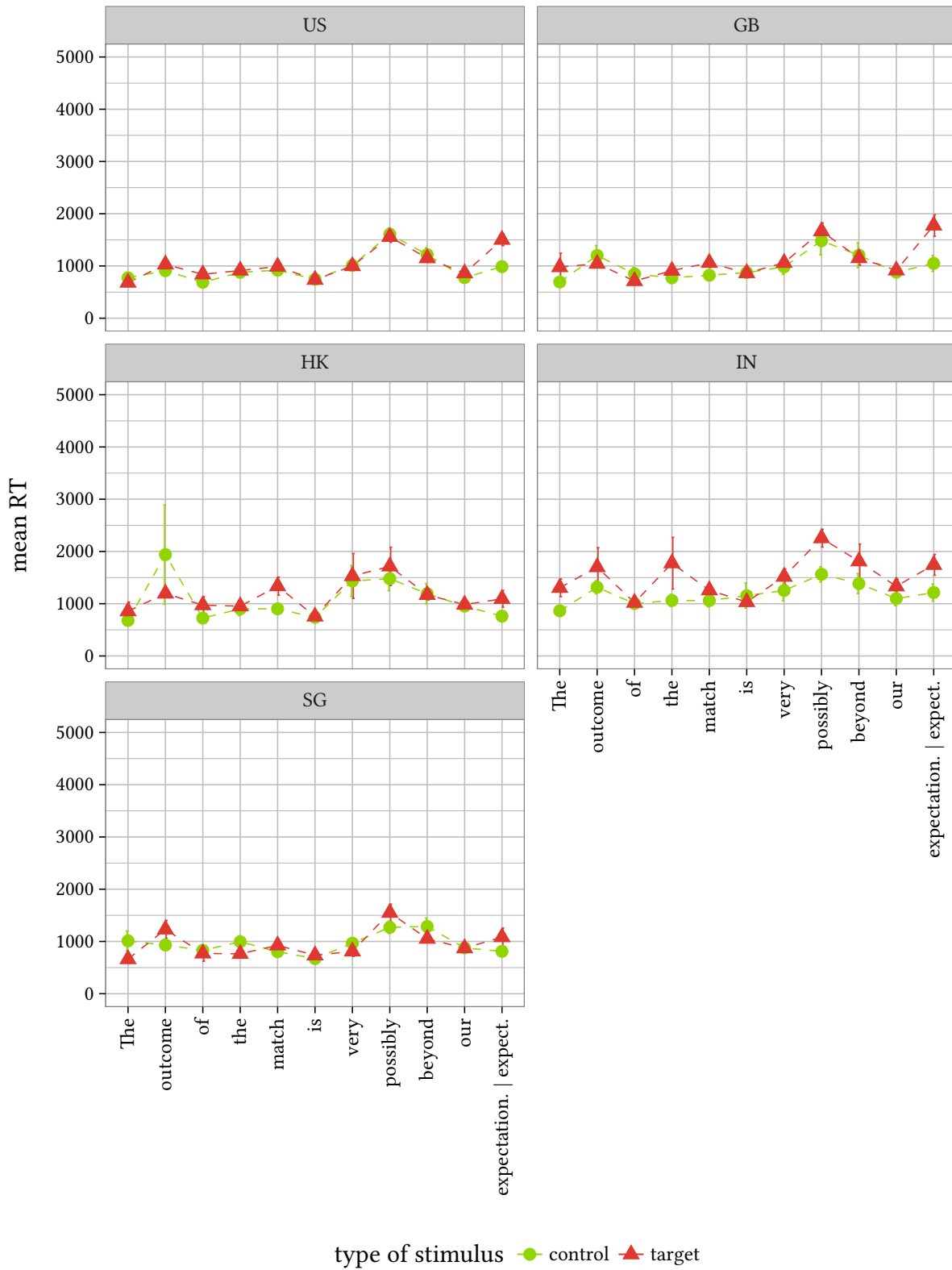


Figure E.13: Reaction times for sentence 21

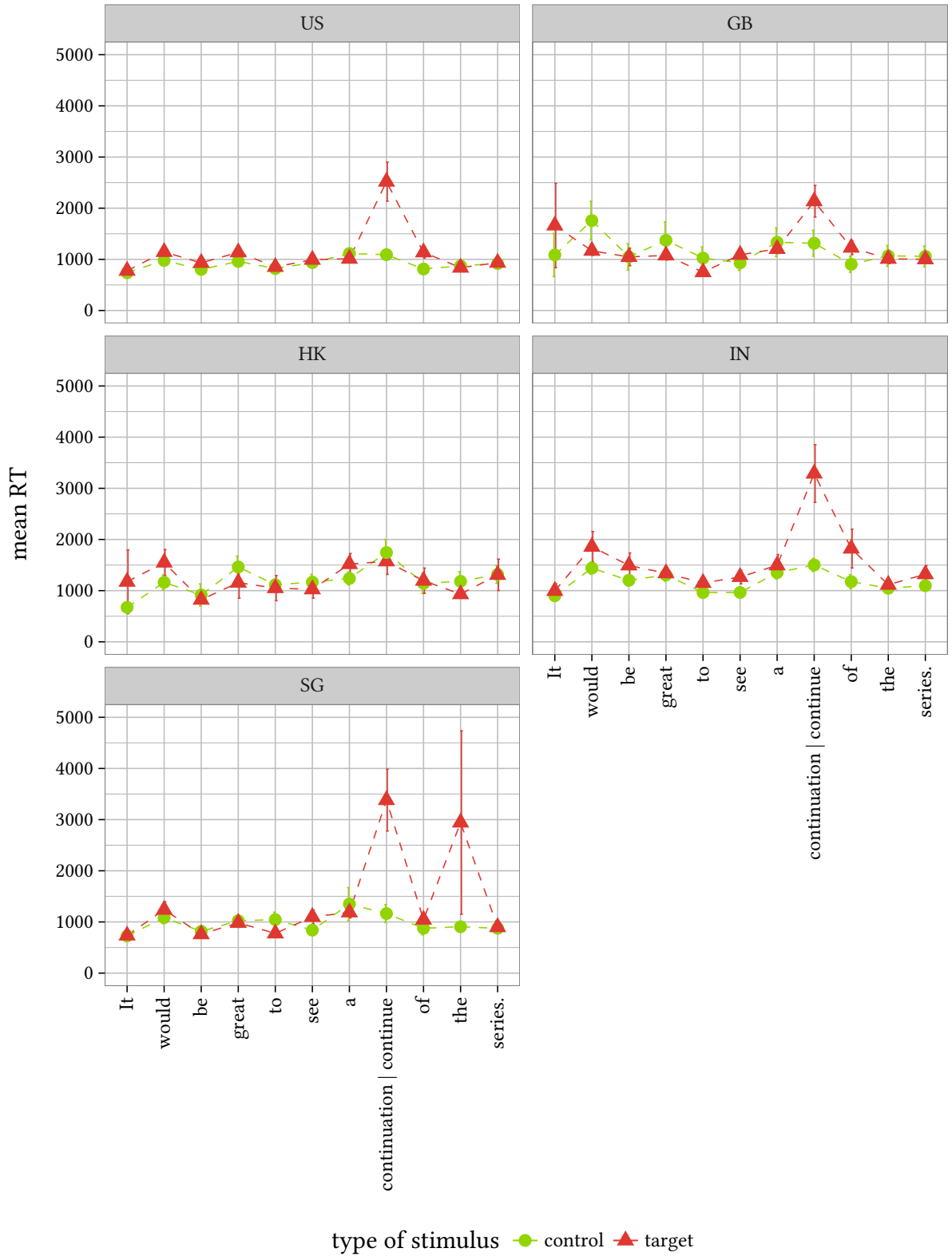


Figure E.14: Reaction times for sentence 25

While phonetic, morphosyntactic, and lexical variation in World Englishes have been studied extensively, phenomena at the lexis-grammar interface have not yet received the same amount of attention. This book investigates the conversion of verbs to nouns as in *to require* (verb) > *a require* (noun) in Asian varieties of English. More specifically, the study compares this process in three New Englishes (Hong Kong English, Singapore English, and Indian English) with its usage pattern in two major native varieties of English: British and US American English.

The methods used to explore this phenomenon range from the quantitative analysis of large corpora such as GloWbE or COCA to the qualitative analysis of the ICE corpora. Corpus findings are subsequently corroborated by means of web-based psycholinguistic experiments testing acceptability as well as processing speed of verb-to-noun conversion.

The main explanatory factors which are scrutinized are, first, the influence of contact languages such as the highly analytic Chinese, second, the degree of institutionalization, which is conceptualized drawing on the developmental phases in Schneider's Dynamic Model, and third, the usage frequencies of the verbal base and the derived, non-converted nominal form (e.g. *requirement*). By applying multivariate statistics, this book provides an attempt to integrate these three factors, contributing to a refinement of the Dynamic Model from the usage-based perspective.

Stephanie Horch studied English and Spanish at the University of Munich (LMU), Germany, and the University of Alberta, Edmonton, Canada. In 2012, she received a teaching degree for secondary education (Gymnasium), and in 2013, an M.A. in English linguistics. She went on to pursue a PhD at the University of Freiburg, Germany, where she was a member of the Research Training Group "Frequency effects in language" (DFG GRK 1624) from 2013 to 2016. This book is a revised version of her dissertation.

ISBN 978-3-928969-68-0



9 783928 969680

UNI
FREIBURG