Morphological simplification in Asian Englishes

Frequency, substratum transfer, and institutionalization

Laura A. M. Terassa

NIHIN studies
New Ideas in Human Interaction

Morphological simplification in Asian Englishes

Frequency, substratum transfer, and institutionalization

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

vorgelegt von

Laura Anna Maria Terassa
aus Nürnberg

Sommersemester 2017

Erstgutachter: Prof. Dr. Dr. h.c. Christian Mair
Zweitgutachter: Prof. Dr. Dr. h.c. Bernd Kortmann

Vorsitzender des Promotionsausschusses
der Gemeinsamen Kommission
der Philologischen und
der Philosophischen Fakultät: Prof. Dr. Joachim Grage

Datum der Disputation: 20.02.2018

Acknowledgements

I am grateful to the many people who have supported me in the course of writing this book. Having already focused on Asian Englishes in my master's studies, I had the opportunity to research a selected number of Asian Englishes from a frequency-based perspective in the course of my PhD project. The project was funded by the German Research Foundation (research training group "DFG RTG 1624: Frequency effects in language").

First and foremost, I would like to thank my supervisors, Christian Mair and Bernd Kortmann, for their continued support and for their constructive feedback. In the three years of working on my project I never lost track or questioned my research agenda, and I am convinced this is to a considerable degree due to the high quality of supervision I enjoyed.

I am grateful to Stefan Pfänder and Christian Mair (speaker and co-speaker of the research training group) as well as to Nikolay Hakimov, Michael Schäfer, and Agnes Schneider-Musah (coordinators; in alphabetical order), for providing us doctoral candidates at the RTG with an inspiring and well-organized environment to work in. Thanks to the professors and to the doctoral candidates affiliated with the research training group for their helpful feedback at seminars and workshops. In particular, I would like to thank Nikolay Hakimov, Adriana Hanuliková, Göran Köber, and Christoph Wolk for their valuable feedback on statistics and experimental design.

Thanks to my office mates Stephanie Horch, Udo Rohe, and Katja Schwemmer (I could not have wished for better ones) and to my colleagues at the English Department, among them Alice Blumenthal-Dramé, Mirka Honkanen, Jakob Leimgruber, and David Lorenz, for their feedback at the *Oberseminar*. In fact, I owe Jakob Leimgruber special thanks because he not only raised my research interest in Asian Englishes but also brought me in contact with the wonderful linguistics team at Nanyang Technological University in Singapore. The team around Ng Bee Chin and Francesco Cavallaro kindly hosted me during research stays in Singapore in 2012 and 2015. I would also like to thank Michael Klotz for (further) raising my interest in linguistics when offering me my first position as a tutor. Many thanks to Bryce Stewart for proofreading the manuscript and to Luke Bradley and Yvonne Chan for

reviewing the stimuli used in the perception experiment. All remaining mistakes are my own.

I was able to present my work at various conferences and workshops on corpus linguistics and World Englishes—among them ChangeE (2015), ICAME (2015, 2016, 2017), and ISLE (2016). I am grateful to the conference organizers for hosting those inspiring events and to the conference attendees for their valuable feedback on my project. Each conference fed me with lots of new ideas. In particular, I would like to mention Bao Zhiming and Susanne Wagner, who gave me helpful input regarding potential substrate influence on nominal plural marking in Singapore English and the use of GloWbE as a frequency database, respectively. Bao Zhiming, Tobias Bernaisch, Elisabeth Bruckmaier, Claudia Lange, and Sven Leuckert spent two days at the University of Freiburg for a workshop entitled "Frequency effects in language contact: Asian Englishes" that I co-organized with two colleagues of mine (Stephanie Horch and Claudia Winkle). Thank you for having been our workshop guests and for discussing our projects and data in great detail.

Heartfelt thanks go out to my parents and my brother for always believing in me and for being there for me unconditionally. Thank you to my close friends for staying close despite the fact that we have not lived in the same region for several years.

Finally, I would like to thank Martin for his never-ending support paired with a healthy portion of pragmatism, for his admirable programming skills, and for his understanding when I spent yet another weekend at my desk. It is him that I dedicate this book to.

Reutlingen, September 2018

Contents

Abbreviations

AIC	Akaike Information Criterion
AmE	American English
AMT	Amazon Mechanical Turk
BNC	British National Corpus
BrE	British English
CSE	Colloquial Singapore English
EFL	English as a foreign language
ENL	English as a native language
ESL	English as a second language
eWAVE	Electronic World Atlas of Varieties of English
FD	Feature density
FRED	Freiburg Corpus of English Dialects
FRIAS	Freiburg Institute for Advanced Studies
GloWbE	Corpus of Global Web-Based English
GloWbE GB	GloWbE Great Britain
GloWbE HK	GloWbE Hong Kong
GloWbE IN	GloWbE India
GloWbE SG	GloWbE Singapore
GloWbE US	GloWbE United States
HKE	Hong Kong English
ICE	International Corpus of English
ICE-GB	ICE Great Britain
ICE-HK	ICE Hong Kong
ICE-IND	ICE India
ICE-SIN	ICE Singapore
IndE	Indian English
IndSAfE	Indian South African English
KWIC	Keyword in Context
LCT	Language Contact Typology

NIECSSE	National Institute of English Corpus of Spoken Singapore English
OED	Oxford English Dictionary
PRC	People's Republic of China
SgE	Singapore English
SLA	Second language acquisition
StE	Standard English
WAVE	Mouton World Atlas of Variation in English

List of figures

List of tables

1 Introduction

> The beginnings of the study of new varieties of English as a serious topic of linguistic research and a new subdiscipline of English linguistics can be dated to the early 1980s, with the publication of some groundbreaking books (Bailey & Görlach 1982; B. B. Kachru 1986, 1992; Pride 1982; Platt et al. 1984; Trudgill & Hannah 1982; Wells 1982) and the launching of scholarly journals devoted to this topic (English World-Wide 1980-, World Englishes 1982-). Prior to that time, no more than a handful of books on some of the major new varieties of English had been published, for example on English in Australia and New Zealand (Baker 1945; Ramson 1966; Turner 1966), West Africa (Spencer 1971), and Singapore (Tongue 1974; Crewe 1977); but there was no overarching awareness of such varieties constituting a joint field of linguistic study, let alone a theory or methodology relating to this topic (Schneider 2003: 234).

The last few decades have seen an increasing body of publications comprising collective volumes (e.g., Schneider 1997 in Schneider 2003: 234), theoretical accounts of the then new area of research (e.g., McArthur 1998 in Schneider 2003: 234), and numerous studies on individual varieties of English. As Schneider (2003) points out, the fact that labels like NEW ENGLISHES, GLOBAL ENGLISHES, and WORLD ENGLISHES are interchangeably used, "is characteristic of a newly emerging field" (234). The more recent debate about the possibly fuzzy boundaries between English as a second language (ESL) varieties and English as a foreign language (EFL) varieties is further proof of a growing field of research whose conceptual boundaries to other fields have aroused research interest. The availability of parallel corpora such as the *International Corpus of English* (ICE; cf. Greenbaum 1996) and the *Corpus of Global Web-Based English* (GloWbE; cf. Davies 2013) in particular has led to a considerable number of studies on single new varieties and variety comparisons. Those studies have improved our understanding of the mechanisms involved in the emergence, development, and change of new varieties. As Mair & Hundt (2000) put

it, corpus linguistics has significantly helped marrying linguistics with "language in the real world" (3), thereby contributing to variety descriptions:

> Over the years [...], corpus linguists have shown that they are interested in detailed and testable accounts of language use in all its baffling complexity rather than a postulated underlying language system, in researching more data on more varieties rather than proposing new analyses for old standard examples, in practical applications of their work in language teaching and translation rather than the ivory tower.

1.1 Simplification

In this book, the focus of attention lies on simplification, a widely discussed phenomenon in the literature on language variation and typology that is related to considerations about degrees of complexity of languages (cf. Kortmann & Szmrecsanyi 2009: 265). In fact, fundamental discussions have evolved over the question whether languages differ in complexity. Proponents of the so-called "equi-complexity axiom," such as Hockett (1958), claim that a trade-off exists that renders all languages (or their grammars, to be more precise) equally complex or simple.

> [I]t would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically (ibid.: 180–181, in Kortmann & Szmrecsanyi 2009: 266).

One of the strong opponents of the "equi-complexity axiom" is McWhorter (2001: 162), who stresses that languages can be grouped along a "scale of complexity" with some languages being more complex than others and pidgins and creoles, in particular, patterning at the end of the scale that marks grammatical simplicity. In line with that, Trudgill (2010: 310–313) points out that pidgins and creoles were originally acquired non-natively by learners beyond the critical threshold in language contact situations. Consider also the following quote (Trudgill 2001: 372):

> Just as complexity increases through time, and survives as the result of the amazing language-learning abilities of the human child, so complexity disappears as a result of the lousy language-learning abilities of

> the human adult. Adult language contact means adult language learning; and adult language learning means simplification, most obviously manifested in a loss of redundancy and irregularity and an increase in transparency. This can indeed be seen at its most extreme in pidgins and hence in creoles [...] But it is not confined to these types of language.

Trudgill (2009) points out that typological differences emerge from different degrees of language contact varieties of English have been exposed to; i.e., high-contact varieties have relatively many grammatically simple features, whereas low-contact varieties are characterized by relatively many grammatically complex features.[1] His distinction between high-contact and low-contact varieties differs from Chambers' (2004) split between vernacular and non-vernacular varieties. According to Chambers (2004), "a small number of phonological and grammatical processes recur in vernaculars wherever they are spoken" (128), which he terms "vernacular universals/roots" (cf. Chambers 2001; 2004). Trudgill (2009) criticizes that the number of vernacular universals is too low to allow for a typological split between vernacular and non-vernacular Englishes though (cf. Kortmann & Szmrecsanyi 2009: 267–268).

The notion of a typological split between high- and low-contact varieties of English also underlies the study on simplification and complexification processes in World Englishes by Kortmann & Szmrecsanyi (2009: 268) just mentioned. The authors compare about 50 mainly non-standard varieties of English with regard to their degrees of morphosyntactic complexity. By numerically quantifying degrees of morphosyntactic simplicity and complexity in the varieties investigated, Kortmann & Szmrecsanyi (2009: 278, 281) show that language contact is likely to reduce complexity levels: Grammaticity levels are highest for traditional L1 vernaculars (e.g., U.S. English, North, Southwest, and Southeast English) and lowest for L2 varieties (e.g., Indian English, Singapore English), while high-contact L1 varieties (e.g., Irish English, African-American Vernacular English) are in between. Grammaticity is defined as "the text frequency of grammatical markers, i.e., their token frequency in naturalistic discourse" (ibid.: 269).

It is worth mentioning here that different approaches to the study of language adopt different complexity measures. While the cross-linguistic perspective of typological research is interested in absolute complexity measures that are theory-driven

[1] For in-depth accounts of simplification and complexification regarding different grammatical features in various languages, the reader is referred to Sampson et al. (2009) and Kortmann & Szmrecsanyi (2012). The contributions therein repeatedly stress the impact of language contact on simplification.

and therefore objective, sociolinguistic and psycholinguistic research works with relative complexity measures that define degrees of complexity from the point of view of the speaker (cf. Szmrecsanyi & Kortmann 2012: 10). Miestamo (2009: 81) puts it as follows:

> The absolute approach defines complexity in objective terms as the number of parts in a system, of connections between different parts, etc. [...] The relative approach to complexity defines complexity in relation to language users: what is costly or difficult to language users (speakers, hearers, language learners) is seen as complex. Complexity is thus identified with cost and difficulty of processing and learning.

It is the latter perspective (the relative complexity measure) that is adopted in this book as well. Thus, it is argued that the features of interest (see section 1.2) are grammatically simpler than their Standard English (StE) counterparts for the speaker groups of interest. Features are likely to be perceived as grammatically simple when they are contact-determined (i.e., when they reflect features speakers know from other languages they speak) or when they are learner features.

1.2 Feature choice

Trudgill (2010: 307–308) mentions loss of redundancy, increased transparency, and regularization of irregularities as typical examples of structural simplification and describes them as "three crucial [...] components to the simplification process" (307). The features of interest in the following chapters are omission and regularization, and they are investigated for both the verbal and the nominal paradigm, as table 1.1 shows. Three Asian Englishes are accounted for, namely Hong Kong English (HKE), Indian English (IndE), and Singapore English (SgE).

Table 1.1: The features of interest by simplification type and paradigm

simplification type	verbal paradigm	nominal paradigm
omission	omission of inflectional past tense marking	omission of inflectional noun plural marking
regularization	regularization of irregular verbs	uncountable nouns treated as countable nouns

Regarding omission, omission of verbal past tense marking and omission of nominal plural marking are considered (compare examples 1.1 and 1.2).[2] The unmarked forms are in italics.

(1.1) Then I *follow* all the <,> all the all the line (ICE-HK:S1A-074#232:1:A)

(1.2) But I just live there with my sisters for almost two years because my *parent* moved (ICE-HK:S1A-014#692:1:A)

Lack of inflectional marking is a clear case of structural simplification because the inflectional suffix is lacking, making the respective verb or noun structurally simpler in that only the base form is left. Of particular interest are cases of lack of inflectional marking where the past time or plural reference is provided in another manner in the sentence or utterance, namely by means of a time adverbial with past time reference or by means of a determiner with plural reference (e.g., a quantifier or a numeral). In those cases, simplification additionally means loss of redundancy in that the twofold (and therefore redundant) marking for past tense or plural (inflectional suffix plus analytic marker) known from StE is dismissed.[3] In the following utterances, the past tense (1.3) and plural (1.4) reference is clear despite the missing inflectional suffix.

(1.3) They chipped in with good discussions and we kind of uh first uh *agree* on uh use of transparencies that is it's something that is a must for a large lectures (ICE-SIN:S2A-047#24:1:B)

(1.4) So I've been here uhm two *year* uh two and half *year* (ICE-HK:S1A-088#87:1:B)

The utterances stem from the Singapore and Hong Kong components of ICE, the former occurring in the ICE section "unscripted speeches," the latter having been uttered in a face-to-face conversation. Table 4.1 in section 4.1 provides an overview of the sections the spoken part of ICE comprises.

Obviously, simplification for the speaker does not (necessarily) result in ease of understanding for the hearer. Thus, while a verb that is inflectionally marked for

[2] For the transcription conventions, see table A.1 in appendix A. The Hong Kong component of ICE will be referred to as ICE-HK, the Indian component as ICE-IND, and the Singapore component as ICE-SIN henceforth. The Great Britain component will be referred to as ICE-GB.

[3] As chapter 5 will show, a distinction needs to be made between formal and functional past tense marking in verbs. Only simple past forms clearly indicate a past time reference, while this is not the case for past participles.

past tense (simple past form only) or a noun that is inflectionally marked for plural clearly provide the past time or plural reference, a verb or noun lacking inflectional marking requires a respective time adverbial, plural determiner (quantifier, numeral, etc.), or the context for the past time or plural reference to be understood.

Less clear cases are nouns and verbs that normally form their past tense and plural in an irregular way and that lack their respective inflectional marking. As the following examples show for verbs, lack of inflectional marking does not (necessarily) make irregular forms structurally simpler:

(1.5) Uh <,> is the marriage <,> when is the marriage I <,> *forget* the date (ICE-IND:S1A-095#187:3:A)

(1.6) You *tell* me before <&> (ICE-HK:S1A-056#117:1:A)

Forget forms its past tense by means of vowel change, *tell* by means of vowel change and affixation. While the respective unmarked forms are clearly simpler from the speaker's (or learner's) point of view in that the base form is used, they are not (necessarily) structurally simpler. The same is true for verbs that form their past tense by means of suppletion. Lack of inflectional plural marking in irregular nouns is not dealt with in this book. The few nouns that form their plural in an irregular way and with sufficient frequencies in ICE to work with (e.g., *child*/*children*, *man*/*men*, *woman*/*women*) would not have made a representative sample (see chapter 6).

Regarding regularization, the regularization of past tense marking in irregular verbs and the use of uncountable nouns like countable nouns are of interest. In contrast with omission of inflectional marking, regularization does not (necessarily) lead to structurally simpler forms. In fact, it is difficult to determine whether *catched* as the regularized past tense form of *catch* is structurally simpler than the irregular form *caught*. Rather, regularization means that irregular forms adopt the major pattern of past tense and plural formation; thus, they are no longer used in an irregular way. Regular forms are simpler than irregular forms insofar as the speaker or learner does not need to recall the respective irregular form but can stick to regular ways of past tense and plural marking, respectively. The following examples are taken from the Hong Kong component of the *Corpus of Global Web-Based English* (GloWbE HK). The regularized forms are in italics:

(1.7) He *leaded* his students to conduct different kinds of experiments to verify and develop new physics theories (GloWbE HK)

(1.8) I get the impression that it's more of a funny cartoon book rather than a non-fiction book where you learn boring *informations* (GloWbE HK)

1.3 Hypotheses

Of particular interest in this book is the interplay of substratum transfer, degree of institutionalization, and usage frequency as potential explanatory factors of omission and regularization. Thus, the book expands the study of World Englishes to the usage-based paradigm. Substratum transfer and degrees of institutionalization of varieties have been of interest in many studies on New Englishes, the latter being certainly promoted by Schneider's (2003; 2007) Dynamic Model that assigns postcolonial varieties of English different degrees of institutionalization depending on a number of linguistic and extralinguistic factors (see below for details). Little attention has been paid to the impact of usage frequency on the development of variety-specific features instead.

Usage frequency

The central claim in usage-based linguistics is that usage patterns shape language acquisition, use, and change (Diessel 2007). Usage patterns are typically measured by means of the frequency with which lemmata, word forms, and features occur. Two important frequency measures that have been widely applied in the field of usage-based linguistics are type and token frequency. While token frequency, or text frequency, describes the absolute number of occurrences of a form in a text, type frequency refers to the number of items that conform to a particular pattern (cf. Bybee 2007: 9).

Despite the fact that the distinction between regular and irregular forms is a traditional one that might imply a stricter division than there actually is, the terminology is commonly applied in the usage-based literature and will also be used here. Irregular verbs are assumed to be stored independently due to the fact that they are not predictable from a schematic representation and are (or used to be) highly frequent in use; i.e. they have or used to have a high token frequency (Croft & Cruse 2004; for details see section 2.2. With regular verbs, it is rather the schematic representation that is entrenched. The regular paradigm is particularly productive because its high type frequency (many verbs form their past tense regularly) makes it easy for verbs to attach the regular past tense suffix.

One of the questions of interest here is whether the different types of entrenchment just described (high token frequency with irregular verbs versus high type frequency with regular verbs) play a role in different degrees of omission in irregular compared to regular verbs. Since irregular verbs are assumed to be highly entrenched due to their high token frequency, it is hypothesized that they are hardly prone to omission of inflectional past tense marking. The same goes for regular verbs because the *-ed* suffix is very prominent in marking the past tense. Should there be a trend for omitting inflectional past tense marking, however, regular verbs are expected to be affected primarily because the individual regular verb in its past tense is not as deeply entrenched as an individual irregular verb is.

The case becomes particularly interesting when regular verbs of different frequencies of occurrence are compared. In line with the previously made assumptions, regular verbs of low frequency are again expected to be more prone to omission than regular verbs of high frequency. For nouns with regular and irregular plural marking, similar ways of entrenchment are likely. Nouns that form their plural in an irregular way will not be considered in this book though. Section 2.2 provides a detailed account of the usage-based paradigm.

In the contact varieties considered, English is mainly acquired as an L2[4], which raises the question whether an observed feature constitutes a learner error or a variety-specific innovation (compare section 2.4). In case omission is a learner feature, infrequent forms are expected to be particularly prone to omission because frequent (marked) target forms are more likely to be entrenched. In case omission is or develops towards a variety-specific innovation, the case is trickier though. Of course, it is reasonable to assume that change (or innovation) is maximally effective when frequent forms are affected (first). However, in line with the Conserving Effect, frequent forms are also particularly well entrenched. If a learner feature develops into a variety-specific innovation, it could well be that this change is initiated by infrequent forms (that are prone to learner errors) but pushed by frequent forms once the "former" learner feature has gained ground. As long as phonetic reduction (Reducing Effect) can be ruled out and omission has not stabilized to a certain degree, it can be assumed that infrequent forms are affected by omission first and frequent forms later. Regular forms are not expected to behave differently from irregular forms in that regard. Hypothesis 1a is as follows:

[4]As we will see, English is increasingly acquired as an L1 in Singapore (section 3.2.1), but it is questionable whether this is also the case for the speakers recorded for the Singapore component of ICE. Metadata to check this is not available.

Hypothesis 1a *In cases where omission of inflectional marking is morphologically conditioned, infrequent forms are affected by omission more (or first) and frequent forms less (or later). This is the case for both regular and irregular forms.*

Let us continue with the regularization features of interest, i.e., the regularization of irregular verbs and the use of uncountable nouns as countable nouns. The regularization of irregular verbs is one of the processes commonly described in the usage-based literature (see section 2.2), whereby irregular verbs whose frequency of use declines regularize because their entrenchment in the mind weakens. The usage-based assumption is that if regularization occurs, it is verbs and nouns of low frequency that are regularized. Highly frequent verbs and nouns are expected to be too entrenched for regularization to occur. Hypothesis 1b is as follows:

Hypothesis 1b *Infrequent irregular forms are affected by regularization more (or first) and frequent irregular forms less (or later).*

Substratum transfer

For the analyses in the chapters to come, varieties were chosen that are either similar or that differ in their substrate languages and degrees of institutionalization. As pointed out before, substratum transfer and institutionalization are investigated here as two further factors besides usage frequency that potentially impact the simplification phenomena of interest. Those factors presumably function as constraints on frequency in case they can explain the observed patterns of simplification while usage frequency cannot.

According to Bao (2010), "[s]ubstratum transfer has attracted, and continues to attract, attention from researchers" (793). In fact, World Englishes research (and research on SgE in particular) has had a strong tradition of focusing on substratum (and superstratum) transfer as a contact phenomenon. Odlin (1989: 12) defines substratum transfer as follows (emphasis in the original):

> *Substratum transfer* is the type of cross-linguistic influence investigated in most studies of second language acquisition; such transfer involves the influence of a source language (typically, the native language of a learner) on the acquisition of a target language, the "second language" regardless of how many languages the learner already knows.

He contrasts substratum transfer with borrowing transfer, which "refers to the influence a second language has on a previously acquired language" (ibid.).

Bao (2017: 312) points out that what transfer exactly is very much depends on the question whether one approaches the transfer process from the perspective of the substrate or the superstrate language. To be more precise, he postulates that "[w]hile substratist theories tend to see substrate influence in terms of transfer, superstratist theories see it in terms of appropriation" (ibid.). In the case of transfer "the said feature transfers from the source to the contact language" (ibid.), while in the case of appropriation "the contact language appropriates the said feature into its own grammar" (ibid.). Similarly, Lefebvre (2015) mentions for creoles that "[t]he contribution of superstrate languages to creoles is mainly discussed in comparison to that of the substrate languages" (177). The contribution of source languages to creole development is little researched.

Trask (2000: 329) raises a critical issue in pointing out that not all irregularities are necessarily explicable by transfer from substrate languages:

> The reality of substrate languages is not in doubt, but many linguists have often abused the idea, attributing every problematic word or form, and every phonological or grammatical change, to the influence of a substrate language about which nothing whatever may be known.

Obviously, substratum transfer (or any other contact phenomenon, for that matter) is not the answer to all irregularities, inconsistencies, or deviations from standard usage. Attempts at explaining features in contact languages by means of substratum transfer or any other language feature need to remain open to the possible finding that language contact may not be responsible for the respective feature development after all.

A first usage-based approach to substratum transfer has been adopted by Bao (2010), who investigates the productivity (i.e., the usage frequency) of transferred features in the target language. Section 3.1, which introduces central approaches towards modeling variation in English, deals with Bao's usage-based account in detail. Bao's account marks a noteworthy step towards combining traditional World Englishes research and usage-based thinking and is therefore worth dealing with in the context of this book.

It is assumed here that in case varieties with similar substratum background are similarly affected by omission and regularization, substratum transfer is likely an explanatory factor. As sections 3.2.2 and 3.3.2 will show, SgE and HKE share a common substratum background. Consequently, in case both varieties reveal similar

omission and regularization patterns that are explicable by substratum influence, substratum transfer is likely to explain the findings.

Language contact creates a multilingual environment speakers are confronted with, where frequency can affect language behavior more, similarly, or less than in pure L1 environments. Here, we go one step further by arguing that substratum transfer is likely to constrain frequency effects in case omission and regularization occur considerably more in SgE and HKE than in IndE irrespective of lemma token frequency (compare hypothesis 2).

Hypothesis 2 *Substratum transfer functions as a constraint on frequency effects in case omission and regularization occur considerably more in SgE and HKE than in IndE irrespective of lemma token frequency, given that the observed simplification patterns can be explained by common substrate influence.*

Thus, in case the omission and regularization patterns observed are explicable by substratum transfer, and in case the phenomena occur in SgE and HKE but considerably less so in IndE, substratum transfer is likely a determinant of simplification. If no clear frequency cline is observable (i.e., forms of different frequencies of use are not affected to different degrees), substratum transfer constrains frequency effects. The same reasoning is applied regarding institutionalizations as a constraint on usage frequency (see below).

Institutionalization

A variety's degree of institutionalization is another factor often dealt with in World Englishes research. This factor has been strongly promoted with the publication of Schneider's (2003; 2007) Dynamic Model, according to which postcolonial Englishes pass certain stages on their way towards becoming independent varieties.

The Dynamic Model draws on four "core parameters" that are related in a monodirectional causal way (Schneider 2014: 11; emphases in the original):

> The sociopolitical and *historical background* in colonial expansion shapes the *identity* constructions of the two main parties involved, the English-speaking settlers in a new region and the locals. These identities (i.e. who feels associated with whom) are decisive for the *sociolinguistic conditions* which shape the communicative settings, and on these, in turn, the resulting *linguistic effects*, the evolving distinctive structural properties of new varieties, are dependent.

On the basis of extralinguistic and (socio-)linguistic conditions (cf. Schneider 2007: 29), the model assigns postcolonial varieties of English to one of five stages: "foundation," "exonormative stabilization," "nativization," "endonormative stabilization," and "differentiation." In stage 1, "foundation," English serves as a mere communication tool between settlers and the indigenous population, two groups that mainly communicate for utilitarian purposes and whose members stay among themselves otherwise. Pidginization and toponymic borrowing take place. In stage 2, "exonormative stabilization," English is adopted in domains such as education, legislation, and administration (ibid.: 34, 36). In this stage, both the settlers and the indigenous population start shifting their identity constructions. While the settlers' English adopts some lexical items of the local variety, the settler variety is considered the linguistics norm, as the term "exonormative stabilization" suggests. Code-switching takes place. Stage 3, "nativization," is characterized by the increasing emergence of local forms due to heavy lexical borrowing. Those who orient towards external norms criticize a deterioration of English though (ibid.: 43). In this stage, the identity gap between the settler strand and the indigenous population is considerably reduced and the settlers feel increasingly at home in the new territory. Stage 4, "endonormative stabilization," is marked by an increasing acceptance of local norms and the local variety of English starts serving as an identity carrier (ibid.: 160). Additionally, after various phases of borrowing and strong changes in phonology, morphology, and syntax, the local variety stabilizes. The settlers consider themselves settled and part of a new nation. In stage 5, "differentiation," the local variety finally stabilizes and allows for internal differentiation (ibid.: 52–55). Similarly, the identity constructions change again. Inhabitants no longer feel as part of a single entity but start forming smaller groups (e.g., by ethnicity, gender).

As we will see, SgE is the most institutionalized variety considered, which is why the most systematic and stable pattern is expected for this variety. Institutionalization functions as a constraint on frequency effects in case omission and regularization patterns are particularly stable in SgE irrespective of lemma token frequency. As regards the impact of a variety's degree of institutionalization on omission and regularization rates, it is assumed that the more institutionalized a variety is, the more stable its omission and regularization rates are. Institutionalization is likely to constrain frequency effects in case omission and regularization patterns are particularly stable in SgE irrespective of lemma token frequency (compare hypothesis 3).

Hypothesis 3 *Institutionalization functions as a constraint on frequency effects in case omission and regularization patterns are particularly stable in SgE irrespective of lemma token frequency.*

Perceiving omission

This book combines corpus analyses with a web-based perception experiment. Investigating perception is a relatively rare endeavor in World Englishes research. For reasons of feasibility, the perception experiment conducted deals with omission of verbal past tense and nominal plural marking exclusively, leaving aside regularization. In the first task, a self-paced reading task, participants read sentences word by word on the screen at their own pace by pressing the space button on their keyboard or tapping the screen on their smart device. Their reaction times were measured. In the second task, an acceptability judgment task, they evaluated sentences with respect to their degree of acceptability on a continuous scale from not acceptable at all to fully acceptable. Chapter 8 presents the experiment design.

The basic assumption underlying the experiment is that lack of verbal past tense and nominal plural marking results in comparatively long reaction times towards the critical (i.e., inflectionally unmarked) verbs and nouns or the words directly following them (self-paced reading task) and in comparatively negative evaluations of the stimuli containing the critical forms, respectively (acceptability judgment task). Speakers of the target varieties (SgE, HKE, IndE) are expected to read inflectionally unmarked verbs and nouns faster and to evaluate them more positively than the control group. The underlying hypotheses are presented in short form, meaning that the different target varieties and features considered are presented in one single hypothesis for each task for reasons of readability.

Hypothesis 4 *Speakers of the target varieties (HKE, IndE, SgE) read verbs that lack inflectional past tense marking and nouns that lack inflectional plural marking compared with the mean of the means of all conditions faster than speakers of the control varieties (AmE, BrE).*[5]

Hypothesis 5 *Speakers of the target varieties (HKE, IndE, SgE) evaluate stimuli containing verbs that lack inflectional past tense marking and stimuli containing nouns that lack inflectional plural marking compared with the mean of the means of all conditions more positively than speakers of the control varieties (AmE, BrE).*[6]

[5] AmE stands for American English; BrE for British English.

[6] The predictor variable "condition" was sum coded to compare the mean reaction time and judgment score for a condition with the mean of the means of all other conditions. This made

As pointed out before, the choice of Asian varieties allows exploration of the impact of substratum transfer and institutionalization on omission rates. Following that reasoning for the experiment, it is assumed that speakers of isolating substrate languages (i.e., speakers of SgE and HKE) are likely to read inflectionally unmarked verbs and nouns faster than IndE and control group speakers. As discussed in section 8.1, it is less straightforward to assume that they also evaluate sentences containing unmarked verbs and nouns more positively though. The reason is that evaluating sentences is not purely performance-based. I.e., the judgments provided do not necessarily represent the initial, spontaneous reaction of participants towards the stimuli (meta-pragmatic assessment). Instead, they may well be influenced by language ideologies or a willingness of participants to please the researcher by guessing what the expected responses are and providing them (observer's paradox).

Regarding the impact of degrees of institutionalization on omission rates, it is assumed that speakers of SgE show more stable reading and judgment patterns than HKE and IndE speakers. As discussed in the development of hypothesis 3 above, SgE is the most institutionalized variety among the three Asian contact varieties considered here and is therefore assumed to be comparatively stable regarding its omission and regularization patterns. This is expected to be true for both the production and perception of verbs and nouns affected by omission.

1.4 Structure of the book

The book is structured as follows: Chapters 2 and 3 further elaborate on the theoretical background the corpus studies and the experiment are based on. Chapter 2 presents simplification as a linguistic phenomenon, including sections on simplification in the *electronic World Atlas of Varieties of English* (eWAVE; cf. Kortmann & Lunkenheimer 2013), on usage frequency as a determinant of simplification, on substratum transfer and institutionalization as constraints on frequency, and on simplification as a learner phenomenon. Chapter 3 presents approaches towards modeling variation in English that are central for this book, and elaborates on the three Asian contact varieties of interest, their degree of institutionalization, and substrate influence on the development of the varieties. Chapter 4 presents the data used and the

more sense than choosing one particular condition as the reference level all other conditions are compared with (see section 8.4 for details).

methods applied and particularly focuses on using the web as a resource in corpus linguistics and on conducting web-based experiments. Chapters 5 and 6 contain the corpus studies on lack of inflectional past tense and nominal plural marking. Apart from a general overview of the corpus findings, the chapters focus on a usage-based account of omission and on the question to which degree substrate influence and institutionalization constrain frequency effects. The corpus studies on regularization are presented in chapter 7, whereas chapter 8 describes the web-based experiment and its findings. While short summaries follow each corpus study and the experiment, chapter 9 takes a broader view and connects the findings on the interplay of substratum transfer, institutionalization, and usage frequency from a bird's eye perspective. Chapter 10 concludes the discussion by placing the findings in the wider context of usage-based linguistics and World Englishes research.

2 More on simplification

This chapter elaborates on simplification in more detail. Section 2.1 focuses on the representation of the features of interest in eWAVE, whereas section 2.2 draws attention to the usage-based paradigm. Section 2.3, briefly recapitulates on two further factors that potentially influence simplification, namely substratum transfer and institutionalization. These factors function as constraints on frequency in case the comparative analyses of the varieties of interest reveal simplification patterns that are explicable by substratum transfer or institutionalization and not exclusively by usage frequency. Finally, section 2.4 investigates how far the features of interest are general learner features and discusses whether the established (but debated) distinction between ESL and EFL can be upheld in that context.

2.1 Simplification in eWAVE

The simplification features of interest for this book were chosen on the basis of the *electronic World Atlas of Varieties of English* (eWAVE; cf. Kortmann & Lunkenheimer 2013), developed at the Freiburg Institute for Advanced Studies (FRIAS) and the English Department of the University of Freiburg, Germany. eWAVE was first released in 2011 as an open access resource, and an updated and extended version has been available since 2013. On its website, eWAVE (Kortmann & Lunkenheimer 2013) is described as follows:

> eWAVE is an interactive database on morphosyntactic variation in spontaneous spoken English mapping 235 features from a dozen domains of grammar in now 50 varieties of English (traditional dialects, high-contact mother tongue Englishes, and indigenized second-language Englishes) and 26 English-based Pidgins and Creoles in eight Anglophone world regions (Africa, Asia, Australia, British Isles, Caribbean, North America, Pacific, and the South Atlantic [...]).

83 leading experts in their respective fields provided ratings for the morphosyntactic features of their varieties. Both descriptive materials and naturalistic corpus

data complete the picture. The *Mouton World Atlas of Variation in English* (WAVE; cf. Kortmann 2012), which was published in early 2013, combines accounts of the variety-specific data sets (including overviews of the (e)WAVE features attested each) with comparisons across variety types and Anglophone world regions. Let us briefly elaborate on the general picture and focus on the location of the varieties of interest therein, before we move on to an account of simplification features in eWAVE. The following summary is based on global analyses of the distribution of all WAVE features by Kortmann & Wolk (2012) but focuses exclusively on the main findings for the varieties of interest. Since eWAVE contains spontaneous spoken English in particular, it is explicitly Colloquial Singapore English (CSE) that is referred to in the catalogue. Among the varieties investigated here, SgE is the only one for which a distinct colloquial variant is described in the catalogue.

As to variety types, Kortmann & Wolk (2012) identify the fewest non-standard features in L2 varieties (the branch CSE, HKE, and IndE cluster in) compared with L1 varieties and pidgins and creoles.[7] It is striking that the L2 varieties outperform the L1 varieties in having fewer non-standard features.[8] For HKE and IndE, the authors report all four top diagnostic features for L2 varieties, which are feature 45 (insertion of *it* where StE favors zero), feature 55 (different count/mass noun distinctions resulting in use of plural for StE singular), feature 100 (leveling of the difference between present perfect and simple past: present perfect for StE simple past), and feature 209 (addition of *to* where StE has a bare infinitive). CSE has three out of the four top diagnostic L2 features (lacking feature 209) and all of the four top diagnostic high-contact L1 features (feature 3: alternative forms/phrases for referential (non-dummy) *it*; feature 66: indefinite article *one/wan*; feature 132: zero past tense forms of regular verbs; feature 174: deletion of auxiliary *be*: before progressive). While CSE has been classified as a high-contact L1 variety in eWAVE, the authors assume that the impact of common local languages makes CSE and HKE

[7] To graphically display the distribution of the morphosyntactic features of interest, the authors use a clustering method that originated in bioinformatics called *NeighborNet* (cf. Bryant & Moulton 2004; Huson & Bryant 2006). By presenting non-hierarchical classifications (cf. Dress & Huson 2004), this clustering method enables "detect[ing] conflicting signals and thus represent[ing] effects of language contact" (Kortmann & Wolk 2012: 919).

[8] On the basis of the features collected in *The World Atlas of Morphosyntactic Variation in English*, Kortmann & Szmrecsanyi (2009) observe the smallest amount of grammatical marking for L2 varieties and conclude that "L2-speakers appear to prefer zero marking over explicit marking, be it (presumably) L2-easy or complex" (278). While this can certainly not explain the comparatively few non-standard features among the L2 varieties collected in the updated eWAVE version, it might hint that L2 varieties do not automatically opt for L2-easy features.

pattern closely together (ibid.: 931). Not included in this Southeast Asian cluster is IndE, which does not closely pattern with any other variety but is placed in between the Southeast Asian cluster and Indian South African English (IndSAfE).

Table 2.1 depicts the L2-simple features among the typically Asian features collected in eWAVE and reveals that there are hardly any. Only features 165 (belonging to the feature area "negation" in eWAVE) and 176 (feature area "agreement" in eWAVE) fall in both feature categories. The bottom half of the table provides the feature description. The typically Asian features listed are those whose feature density (FD Asia) is 20 percent higher than the respective global feature density (FD world; cf. Mesthrie 2012: 786). Mesthrie (2012) uses "Edgar Schneider's criterion of 80% as a cut-off point to indicate feature density, i.e., that a particular feature occurs in 80% of the Asian varieties categorized in WAVE" (786). The classification of the features according to L2 acquisition difficulty (i.e., as L2-simple features) was adopted from Szmrecsanyi & Kortmann (2009: 69–71). The authors define L2 acquisition difficulty as "the degree to which a given variety does *not* attest phenomena that L2 acquisition research has shown to recur in interlanguage varieties" (ibid.; emphasis in the original).

Table 2.1: eWAVE L2-simple features among features that occur more frequently in Asia than globally (by feature density (FD); Mesthrie 2012)

feature no.	FD Asia (%)	FD world (%)	difference (%)	feature area
165	100.0	66.2	33.8	negation
176	71.4	39.2	32.2	agreement

feature no.	feature
165	invariant non-concord tags
176	deletion of copula *be*: before NPs

Table 2.2 provides an overview of the features chosen for in-depth investigation here and the respective ratings for the contact varieties of interest. Information on whether the feature is one of the top L2 features (cf. Kortmann & Wolk 2012), one of the top Asian features (cf. Mesthrie 2012), or one of the L2-simple features (cf. Szmrecsanyi & Kortmann 2009: 69–71; Kortmann & Szmrecsanyi 2009: 274) is also included. Again, the bottom half of the table contains the feature description.

With regard to the ratings, it is worth noting that all features except 55 and 56 have C-ratings in IndE but A- or B-ratings in CSE and HKE. Feature 56 arguably

Table 2.2: The features of interest in eWAVE (Kortmann & Lunkenheimer 2013)

feature no.	ratings†			top L2	top Asia	L2-simple
	CSE	HKE	IndE			
55	B	A	A	✓	✓	
56	D	D	B			(✓)
57	A	A	C		✓	
58	A	A	C		✓	
128	B	A	C			✓
132	A	A	C			✓

feature no.	feature
55	different count/mass noun distinctions: use of plural for StE singular
56	absence of plural marking only after quantifiers
57	plural marking generally optional: for nouns with human referents
58	plural marking gen. optional: for nouns with non-human referents
128	regularization of irregular verb paradigms
132	zero past tense forms of regular verbs

†ratings in eWAVE: A - feature is pervasive or obligatory; B - feature is neither pervasive nor extremely rare; C - feature exists, but is extremely rare; D - attested absence of feature; X - feature is not applicable; ? - no information on feature is available

levels out features 57 and 58 because plural marking is classified as generally optional in CSE and HKE, irrespective of preceding quantifiers. Feature 55 is the only feature that is among the top L2 features, and features 57 and 58 occur with considerably higher feature density in the Asian varieties in eWAVE (top Asia) than globally (i.e., across all eWAVE varieties irrespective of world region). The features of interest in this book regarding the verb phrase (regularization of irregular verbs, lack of inflectional past tense marking) have been classified as features that ease L2 acquisition (L2-simple), the features of interest regarding the noun phrase (uncountable nouns treated as countable nouns, lack of inflectional noun plural marking) not; except for feature 56. In the first (e)WAVE version, which the categorization according to L2 acquisition difficulty undertaken by Szmrecsanyi & Kortmann (2009: 69–71) is based on, feature 14 (absence of plural marking after measure nouns) was listed, which is not synonymous with feature 56 in the updated eWAVE version. Thus, the check mark for feature 56 as an L2-simple feature is given in brackets. In sum, the features of interest are a mixed bag of top L2, top Asian, and L2-simple features.

As table 2.1 revealed, among the top Asian features there are only two features (156 and 176) that are considered to be L2 features in Szmrecsanyi & Kortmann (2009), so it was decided to take both the top Asian and L2-simple features into account and not necessarily combinations of the two.

The classification of L2-simple features by Szmrecsanyi & Kortmann (2009: 69–71) provides a general overview of simplification in the varieties of interest—even more so as it seems to be the first and only systematic account of simplification across varieties of English. Despite the fact that SgE is increasingly developing towards an L1, CSE has three out of the four top diagnostic L2 features in eWAVE and patterns closely with HKE (a typical L2). Simplification operates in many different ways and since any comprehensive approach towards simplification is necessarily a simplification itself, the focus on L2-simple features seems justified to get an overall idea of simplification in the varieties considered here.

2.2 Usage frequency as a determinant of simplification

The usage-based approach to the study of language acquisition, use, and change developed as a response to Chomsky's "rigid division between grammar and language use" (Diessel 2007: 108), i.e., the clear distinction between competence and performance. This fundamental distinction dates back to Ferdinand de Saussure (1916), according to whom there is a difference between the knowledge speakers have about their language and the way they use their language (cf. Bybee 2007: 6). American structuralists in general and generativists in particular postulate that only the knowledge of language is worth studying, whereas frequencies of use do not impact grammatical structures and are therefore negligible (cf. Chomsky 1965 and later works; Bybee 2007: 6).

The generativist view has been challenged by functionally oriented linguists and scholars in cognitive science who argue that usage frequencies fundamentally influence language structure and use (cf. Diessel 2007: 109). In recent decades[9], the argument has been underpinned by evidence from studies on language acquisition, grammaticalization, and linguistic typology, among others (e.g., Barlow & Kemmer

[9]Bybee (2007: 6–7) mentions two reasons why the functionalist approach to the role of frequency lacked empirical research in the beginning. Firstly, theoretical and methodological gaps between linguistics and psycholinguistics made theory building on the basis of psycholinguistic findings difficult. Secondly, the lack of empirical proof of "mental representations" postulated by generativists made "the more empirically minded functionalists" (ibid.: 7) hesitate to formulate assumptions about effects of repetition on the mind.

2000; Hawkins 2004), and the impact of frequency on language production, comprehension, and the development of linguistic categories has been stressed (e.g., Bod et al. 2003). Bybee (2006: 711), whose research has largely shaped usage-based theory, emphasizes the importance of experience with language (in the sense of frequency of occurrence and repetition) for rule formation and language change as follows:

> [T]he general cognitive capabilities of the human brain, which allow it to categorize and sort for identity, similarity, and difference, go to work on the language events a person encounters, categorizing and entering in memory these experiences. The result is a cognitive representation that can be called a grammar. This grammar, while it may be abstract, since all cognitive categories are, is strongly tied to the experience that a speaker has had with language.

Thus, in contrast with generativist accounts of language, Bybee (2006: 711) points out that experiences a person has with language significantly influence the cognitive representation (i.e., the grammar) in that person's mind and are therefore of great importance for the study of language use, development, and change. The following paragraphs repeatedly refer to Bybee's usage-based account to language and to related work that underlies her arguments.

Crucial for usage-based theory is the assumption that general knowledge can be derived from item-specific knowledge without item-specific knowledge getting lost in the process. This has been known in psychology since the late 1960s and the 1970s, when researchers showed that test subjects were able to assign stimuli (colored geometrical objects, dots, and lines arranged in ways to represent facial features) to categories on the basis of the similarity of the stimuli to a prototype the subjects had not been presented with (cf. Posner & Keele 1968 and Medin & Schaffer 1978, in Bybee 2006: 8). The behavior of the test subjects reflects both item-specific as well as general knowledge about the stimuli. Additionally, it was proven that repetitions impact category formation: Different stimuli that were shown repeatedly led to relatively high sensitivity to individual patterns, similar stimuli shown only once each led to relatively high sensitivity to the prototype tendency.

Similarly, frequency or repetition is important for language development and change because mental representations of language are influenced by repetition (cf. Haiman 1994, in Bybee 2007: 8). Two seemingly contradictory major findings noted by Hooper (1976) are worth mentioning here and will be elaborated on in the following paragraphs: Firstly, regularization (a form of analogy) occurs first in infrequent

paradigms and only later in highly frequent paradigms. Secondly, highly frequent words are prone to sound change first, infrequent words only later.

As mentioned in section 1.3, two important frequency measures that have been widely applied in the field of usage-based linguistics are type and token frequency. The distinction between type and token frequency can be nicely explained by means of the development of patterns of past tense marking in English, which is also a prime example for explaining the abovementioned finding that regularization occurs first in infrequent paradigms and only later in highly frequent ones. The historical development of English verbs shows that the majority of verbs have undergone a regularization process throughout time and form their past tense by means of adding the affix {*-ed*} today (ibid.: 10). Nevertheless, a group of mostly highly frequent irregular verbs has survived until today in that the verbs have not adopted the major morphological process of past tense formation. Commonly known as irregular verbs, these verbs form their past tense by means of suppletion (e.g., *go/went/gone*), vowel change (e.g., *read/read/read*), or a combination of vowel change and affixation (e.g., *bring/brought/brought*) instead. Bybee (2007: 10) explains this phenomenon with the so-called Conserving Effect. The high frequency of irregular verbs has strengthened their memory representation, which is why they have resisted the pressure to adopt affixation of the morpheme {*-ed*} for marking past tense. The strengthened memory representation of highly frequent forms also implies that those forms are particularly easily accessible, i.e., they can be retrieved from the mind relatively easily (ibid.: 10).

Since a relatively small number of verbs form their past tense by means of suppletion, vowel change, or vowel change plus affixation, the irregular pattern of past tense formation has a lower type frequency than the regular pattern. The regular pattern is more productive in that its schematic representation is stored and can be "easily combined with a particular [verb] to form [the past tense]" (Croft & Cruse 2004: 296). With its three allomorphs (/t,d,ɪd/), the past tense schema has for each allomorph "a high type frequency of low token frequency instances, so each allomorph is highly productive for its phonologically defined class" (ibid.: 296). Due to the complementary phonological distribution of the allomorphs based on the final sound of the verb stem and their identical meaning (past time reference), speakers may derive a superordinate category [*ed*/PAST] (ibid.: 297).

The distinction between regular and irregular verbs is a very traditional one (Croft & Cruse 2004: 292). According to traditional models of language, there is a straight-

forward structural relationship in regular verbs in that the past tense of regular verbs can be directly formed from the bare verb lemma by adding the *-ed* suffix to the base. This is not the case for irregular verbs, which are "listed in the lexicon" (ibid.) as such. Croft & Cruse (2004) point out that word forms can also be entrenched[10] in case they are "predictable from a more schematic representation" (ibid.: 292). Thus, the comparatively frequent plural form *boys* is more likely to be entrenched than the infrequent plural form *cornices*, although both plural forms are examples of regular plural formation. The authors hypothesize that "the storage of a word form, regular or irregular, is a function of its token frequency" (ibid.).

Irregular forms are assumed to be stored independently due to their lack of predictability from a schematic representation. When their frequency of use declines, they are predicted to become regularized. Bybee & Slobin (1982: 270) show, for instance, that preschool children are particularly likely to regularize irregular verbs of low token frequency. Production experiments with children in the third grade and adults revealed similar results (ibid.: 270–271). For verbs with regular past tense marking, Stemberger & MacWhinney (1988: 106) observe that participants in an experiment who had to produce the respective past tense forms at high speed made significantly less mistakes with regular verbs of high frequency than with regular verbs of low frequency.

Seemingly contradictory (cf. Hooper 1976), it has been shown that highly frequent words are prone to sound change first (and at a faster rate), whereas infrequent words undergo sound change only later (and at a slower rate; early contribution by Schuchardt 1885, in Bybee 2007: 235; Bybee 2000 on AmE t/d deletion; Jurafsky et al. 2001 on the effect of word predictability on articulatory reduction).[11] Well-known examples of this so-called Reducing Effect are phonetically reduced greetings such as *God be with you* shortened to *goodbye* (cf. Bybee 2007: 11). Across grammatical items, "repetition of neuromotor sequences" (ibid.: 11) results in increased overlap and reduction of the respective articulatory gestures. Consequently, highly frequent words are reduced first. According to Bybee, neuromotor sequences are processed as single units (ibid.). The gradual spread (of reduction) from highly frequent to

[10] Langacker (1987: 59–60) uses the term "entrenchment" to describe the independent storage of frequent word forms.

[11] Schuchardt (1885) was the first to claim that sound change does not occur regularly in a way that phonological differences across languages or dialects affect all words with the respective phonetic environments equally (cf. Bybee 2007: 236). Compare also Wang (1969; 1977), Wang & C.-C. Cheng (1977), Labov (1981; 1994), and Phillips (1984), among others, for similar accounts.

infrequent forms is known as lexical diffusion. Zipf's (1935) explanation as to why high frequency tends to occur with reduced (and therefore shorter) forms is also worth mentioning in that context. According to him, abbreviations (e.g., *laboratory* turning into *lab*) help retain a balanced frequency distribution of long and short words in languages (cf. Bybee 2007: 12).

Chunking (i.e., the transfer of sequences to a higher level where they comprise a single unit) is another effect resulting from the repetition of sequences. While chunking is particularly common in highly frequent sequences (e.g., *I don't know*), Jurafsky et al. (2001), among others, provide evidence that reduction between words depends on the frequency with which the words co-occur. Thus, besides the individual frequencies, it is the interdependence of the words that add up to a chunk that is of importance.

As the previous paragraphs have shown, the seemingly contradictory results of the Conserving Effect (which discourages change) and the Reducing Effect (which encourages change) are based on different types of change that take place. While frequent and complex units are unlikely to change due to their strengthened memory representation (Conserving Effect), often repeated units (e.g., chunks) are particularly prone to sound change for articulatory reasons (Reducing Effect). In the previous paragraphs, the reader will have noticed that some of the features of interest here have been extensively discussed in the usage-based literature just presented.

2.3 Substratum transfer and institutionalization as constraints on usage frequency

Of particular interest in this book is the question of how far substratum transfer and institutionalization constrain frequency effects. As pointed out before, substratum transfer and institutionalization are factors traditionally accounted for in World Englishes research. Usage-based reasoning has had little impact on variety descriptions instead. In combining usage frequency, substratum transfer, and institutionalization as potential determinants of simplification, this book explores the limits substratum transfer and institutionalization set for usage frequency.

In section 1.3, it was hypothesized that substratum transfer functions as a constraint on frequency effects in case omission and regularization occur considerably more in SgE and HKE than in IndE irrespective of lemma token frequency. Obviously, the observed simplification patterns need to be explicable by common sub-

strate influence for the argument to hold. Similarly, it was hypothesized that institutionalization functions as a constraint on frequency effects in case omission and regularization patterns are particularly stable in SgE irrespective of lemma token frequency.

A constraining effect of substratum transfer and institutionalization on usage frequency is most clearly visible in case omission and regularization prevail across verbs and nouns, whether the verbs and nouns are regular (and therefore rather infrequent) or irregular (and therefore rather frequent). Thus, given that omission and regularization are explicable by substrate influence, substratum transfer constrains frequency effects in case regular and irregular verbs and nouns are affected by omission and regularization to similar degrees. Leaving the impact of substratum transfer aside, usage-based reasoning per se would imply that frequent and infrequent forms are affected by omission and regularization to different degrees. Similarly, given omission and regularization patterns are particularly stable in SgE, the most institutionalized variety accounted for here, institutionalization constrains usage frequency in case omission and regularization patterns in frequent and infrequent verbs and nouns are similarly stable in SgE and less stable (across frequencies) in HKE and IndE.

As those assumptions show, the variety choice enables disentangling to which degree substratum transfer and institutionalization account for the observed omission and regularization patterns. Should SgE and HKE behave similarly and should the observed patterns be explicable by substratum transfer, substratum transfer is likely an explanatory factor. Similarly, should omission and regularization patterns in SgE be particularly stable (compared with those in HKE and IndE), the high degree of institutionalization of SgE is likely to account for the respective degree of feature stability. How substratum transfer and institutionalization affect the simplification features of interest will be dealt with in the respective analysis chapters (chapter 5 for omission of verbal past tense marking, chapter 6 for omission of nominal plural marking, and chapter 7 for the regularization of irregular verbs and the use of uncountable nouns like countable nouns). As mentioned above and as the analysis chapters will show, substratum transfer is typically accounted for in the literature on contact varieties, but to varying degrees. Substratum influence on SgE has been extensively described, for HKE the contrary is the case.

2.4 Simplification as a learner phenomenon

The simplification features considered here are investigated for two L2 varieties (HKE, IndE) and one former L2 variety that is increasingly developing into an L1 (SgE). One immediate question that emerges is whether simplification in the varieties of interest is the result of innovative language use or rather an expected phenomenon (perhaps even an error) in languages acquired by speakers of different mother tongues. As Edwards (2014) points out with reference to Biewer (2011) and Schneider (2012), among others, "[i]t has repeatedly been noted that [ESL and EFL varieties] share a common acquisitional starting point, which results in similar strategies such as transfer, redundancy and regularisation" (Edwards 2014: 173).

There is, in fact, a recent and growing body of literature on the distinction between ESL and EFL learners and on the deviations of their Englishes from standard usage (typically referred to as "innovations" in the case of ESL and as "errors" in the case of EFL). The debate was inspired by Sridhar and Sridhar's (1986) criticism of a "paradigm gap" between research on learner Englishes (following the tradition of second language acquisition (SLA) research) and research on institutionalized L2 varieties of English and their "plea for an integrated approach" (Hundt & Mukherjee 2011: 1). Traditionally, the term "second language variety" or "L2" has been used for varieties that are mainly acquired as a second language by most members of a speech community. This is typically the case for Outer Circle countries in Kachru's (1985; 1988; 1992) Three Circles model. The term "learner English" is linked to Selinker's notion of interlanguage and refers to the linguistic mental system learners of a second language develop upon acquiring a second language. It takes account of the "idiolect of individual speakers" (Bongartz & Buschfeld 2011: 37) rather than describing the status of a variety in a speech community in general. Learner Englishes usually have foreign language status in Expanding Circle countries in Kachru's model (ibid.).

Despite the ongoing debate, up until today few studies have "explicitly compare[d] L2 and learner varieties" (Edwards 2016: 4). Edwards (2016) mentions Nesselhauf (2009) who observes clear parallels in new prepositional verbs emerging in both L2 and learner Englishes as one of the few exceptions. Edwards (2016) points out that "[s]uch findings highlight the paradox that the 'innovations' identified in ESL varieties tend to coincide with those held up as common 'errors' in EFL" (4).

Bamgbose (1998) raises the issue that "with innovations [there is] the need to decide when an observed feature of language is used indeed as an innovation and

when it is simply an error" (2). According to him, innovations differ from errors in that they are widely used in a speech community, increasingly institutionalized (e.g., accepted by authorities, used in textbooks), and commonly accepted among speakers.[12] Crucially, Bamgbose (1998) points out that "[i]nnovations in non-native Englishes are often judged not for what they are or their function within the varieties in which they occur, but rather according to how they stand in relation to the norms of native Englishes" (1). This, however, means that any judgment of innovations according to external norms necessarily takes native Englishes as the varieties the innovated forms are compared with.

Van Rooy (2011) argues for two criteria that allow identification of "conventionalized innovations," as he calls them: grammatical stability and acceptability. According to him, "[t]he distinction between error and conventionalized innovation is one of the crucial issues that researchers dealing with New Varieties struggle to come to terms with, and a key area in which more progressive views of 'varieties' are open to criticism" (ibid.: 191). One exception he mentions is Schneider (2007: 97–109), who "posit[s] structural nativization as a central contributor to the reshaping of English" (Van Rooy 2011: 191), and thus explicitly acknowledges "the possibility that non-native performance phenomena may well feed into the linguistic feature pool" (ibid.). In fact, Schneider (2007: 102–107) refrains from calling processes like restructuring or simplification "errors." For Van Rooy (2011), "[t]he crucial issue to resolve is whether the New Englishes exhibit genuine linguistic innovations that become conventionalized, or whether they simply exhibit errors" (192). He refers to Croft's (2000) distinction between linguistic innovation and linguistic conventionalization, linguistic innovation meaning the individual (usually unintentional) creation of new forms in communication. Innovations (not seldom in the form of performance errors due to transfer, simplification, etc.) particularly occur in language contact settings and can finally become conventionalized and entrenched in case they diffuse and are selected into the linguistic system of the speech community by means of social forces.[13] On that basis, Van Rooy (2011: 192–193) reconceptualizes the difference between errors and conventionalization insofar as errors are common among both speakers of Foreign Language Englishes and speakers of New Englishes, whereas

[12] Bamgbose (1998: 3) distinguishes five internal factors that define the status of an innovation, namely the demographic factor, the geographical factor, the authoritative factor, codification, and acceptability. For detailed accounts of these factors, compare Bamgbose (1998: 3–5).

[13] Van Rooy (2011) extends the social forces operating on linguistic diffusion in contact settings to New Englishes.

conventionalizations are more likely in New Englishes than in Foreign Language Englishes. While both types of varieties have a similar starting point because their learners acquire them as an L2, the social context makes the difference: Speakers of New Englishes come in contact with English on a much more regular basis than speakers of Foreign Language Englishes do because English has a number of internal functions in education, administration, and the like in New Englishes settings. Additionally, New Englishes are likely to serve as the means of communication among speakers of different language backgrounds, and the norm orientation is considerably lower than with Foreign Language Englishes that are mostly acquired in an educational context. Van Rooy (2011: 194) argues that the identity dimension in Schneider's (2007) Dynamic Model plays a crucial role in that former "errors" can find their way into the speech community once the indigenous population is accepted by the settlers as part of a common speech community.

As we have seen, a debate has evolved around the distinction (or lack thereof) between ESL and EFL and, connected with that, the decision whether to speak of "innovations" or "errors" when deviations from native Englishes are encountered. It is certainly the latter distinction (innovations versus errors) that is of particular importance here. Van Rooy's (2011) criteria to identify conventionalized innovations (namely grammatical stability and acceptability) that set conventionalizations apart from learner errors are a handy tool for identifying what the simplification features of interest actually constitute. With regard to the ESL-EFL distinction or continuum, it is worth pointing out that indeed the question has been raised whether HKE is acquired as a second or as a foreign language. As section 3.3 will show, the fact that English is one of the official languages of Hong Kong and is gaining importance apart from that can be interpreted as a sign of its L2 status (positioned closer towards learner Englishes than SgE, if an ESL-EFL continuum is assumed). Obviously, the linguistic reality does not (necessarily) reflect political decisions on uses of a language, but census data (that admittedly rely on self-reports) show tendencies that should not be neglected (see section 3.3.2).

3 The Asian Englishes of interest

This chapter draws attention to the three Asian varieties of interest. Before turning to the varieties themselves, it deals with approaches towards modeling variation in English that are central in the context of this book. For each variety, the chapter then provides a concise introduction to the colonial and postcolonial history of the setting and the society the variety has emerged in, including information on the variety's institutionalization path and current patterns of language use that are approximated both from the literature and by means of recent census data. Additionally, the chapter elaborates on substrate influence the Asian Englishes of interest have been and are exposed to.

3.1 Modeling variation in English

In the last few decades, a number of models have been developed that categorize varieties of English around the world. This section focuses on central models dealt with in this book and in doing so presents the "Language Contact Typology (LCT) of World Englishes" recently developed by Onysko (2016).

Kachru's (1985; 1988; 1992) Three Circles model

One of the most influential models on World Englishes is Kachru's (1985; 1988; 1992) Three Circles model. Kachru's own visualization of the model consists of a number of oval shapes, whereby the lowest (empty) circles stand for early stages in the history of English, such as Old English and Middle English, and the upper circles represent Inner Circle varieties, Outer Circle varieties, and Expanding Circle varieties (from bottom to top; B. B. Kachru 1992). Later visualizations depict those three variety types by means of three concentric circles. The distinction between Inner Circle varieties, Outer Circle varieties, and Expanding Circle varieties is essentially the same as the distinction between ENL (English as a native language) varieties, ESL varieties, and EFL varieties. Inner Circle varieties are commonly acquired as native languages, whereas the Outer Circle comprises varieties that typically function as official languages and that are likely to be acquired early as second languages

(ibid.). Outer Circle varieties are described as having nativized registers and "an extended range of use" (B. B. Kachru 1985, in Meierkord 2012: 4). Expanding Circle varieties, finally, "have a highly restricted functional range in specific contexts; for example those of tourism, commerce, and other international transactions" (ibid.). In Expanding Circle countries, English is taught as a foreign language.

Kachru's (1985; 1988; 1992) Three Circles model adopts a nation-based perspective of the Englishes represented without accounting for variety-internal variation or the influence of sociolinguistic factors (cf. Bruthiaux 2003). Nevertheless, the model has pioneered the conceptualization of the spread of English around the world in that it shifted the attention away from a "monolithic view of English" (Onysko 2016: 199) towards an acknowledgment of various Englishes and their differences in development and use (cf. Bolton 2006: 241).

Schneider's (2003; 2007) Dynamic Model

A highly influential model that has largely shaped recent discussions of the status of World Englishes is Schneider's (2003; 2007) Dynamic Model, which has already been elaborated on in the introductory chapter (section 1.3). The model is used in this book to approximate degrees of institutionalization. As mentioned in section 1.3, Schneider's Dynamic Model assigns postcolonial Englishes to different stages in their development towards institutionalized varieties. The stages are called "foundation," "exonormative stabilization," "nativization," "endonormative stabilization," and "differentiation." For details on central characteristics of those stages, the reader is referred to section 1.3.

In a reflection of his Dynamic Model and research thereupon, Schneider (2014) addresses the question whether the Dynamic Model, which "focuses explicitly on 'Postcolonial' Englishes (of the Inner and Outer Circles)" (10), is applicable to Expanding Circle varieties as well. He investigates whether stage 2 to 4 components of the Dynamic Model are present, absent, or of unclear status in a number of Englishes in the East Asian Expanding Circle (English in China, (South) Korea, Japan, and ASEAN).[14] What Schneider (2014) observes "is distantly related to what the Dynamic Model describes [but] works only on a rather general level, with some degree of abstraction" (27). In order to account for the underlying dynamisms, Schneider (2014) proposes the term "transnational attraction," which he defines as "the appropriation of (components of) English(es) for whatever communicative purposes

[14]The question whether one can actually refer to the respective Englishes by means of regionally specific terms (e.g., China English) is not addressed.

at hand, unbounded by distinctions of norms, nations or varieties" (28). English is referred to as "an economic resource" (B. B. Kachru 2005: 91, in Schneider 2014: 28) the use of which is shaped by "utilitarian considerations" (Schneider 2014: 28).

While Onysko (2016) approves of the "inclusion of exo- and endonormative forces, nativization, and differentation [as] relevant aspects in the (post)colonial contact-development of Englishes in the world" (201), he criticizes that the Dynamic Model suggests a linear development of postcolonial Englishes. In that context, he mentions Evans' (2014) application of the Dynamic Model to HKE, an Outer Circle variety in Kachru's (1985; 1988; 1992) Three Circles model. Rather than relying on "the synthesis of synchronic evidence from assorted secondary sources" (Evans 2014: 595), Evans examines "evidence from corpora of Legislative Council proceedings and English-language newspapers, census reports, jury lists together with material from private papers, government reports and Colonial Office correspondence" (ibid.) to test the stages HKE went through according to Schneider's (2003; 2007) Dynamic Model. To mention just one central observation, Evans (2014) points out that English-knowing bilingualism in Hong Kong in the late twentieth century resulted from "[t]he shift from elite to mass English-medium schooling in the 1970s and 1980s" (596). Increasing interactions between members of the indigenous strand (consisting of settlers, transients, and refugees in the beginning) and members of the settlers' strand (typically sojourners), which the Dynamic Model accounts for, contributed little to the development of HKE instead (ibid.). In fact, Evans (2014) questions the applicability of the model to Outer Circle varieties in general because "to date, research into these varieties has centred on documenting their linguistic features (e.g., Mesthrie 2008) rather than on collecting baseline sociolinguistic data about their users and uses, both past and present" (Evans 2014: 596). For a more detailed account of Evans (2014), the reader is referred to section 3.3, which deals with the application of the Dynamic Model to HKE.

Onysko's (2016) Language Contact Typology (LCT) of World Englishes

In a recent response to previous attempts at modeling variation in World Englishes (see above), Onysko (2016) proposes the Language Contact Typology (LCT) of World Englishes. The LCT draws attention to the role language contact plays in the development of varieties of English, "offers a reframing of our perspective on world Englishes" (ibid.: 197), and "complement[s] other models in the research paradigm" (ibid.). The underlying assumption is that language contact is a "basic principle of language development that [. . .] underlies all types of Englishes" (ibid.).

The LCT offers a "multidimensional view of language contact" (ibid.: 215) that considers settings, processes, and parameters of contact, and that groups World Englishes into five categories. The LCT follows psycholinguistic research on bi- or multilingual speakers such as Dijkstra & Van Heuven (2002) and Finkbeiner et al. (2006) in that it addresses "the actual cognitive events of contact" (ibid.: 209) by conceptualizing "language as activity in a neuronal network that engages all cognitive capacities of a human being" (ibid.). Onysko (2016: 209) mentions the cultures involved, the directionality, duration, and history of the contact, the medium (e.g., direct speaker interaction) and mode (written/spoken) of the contact, and the way in which the contact is embedded (role of institutions, for instance) as factors that define every contact situation.

Regarding the processes underlying language contact, Onysko (2016) presents "three different cognitive (and systemic) pathways [that] give rise to different, observable contact phenomena" (209–210). They are depicted in table 3.1.

Table 3.1: Processes of language contact and observable contact phenomena (adopted from Onysko 2016: 210)

cognitive/systemic processes		observable contact phenomena
co-active selection of linguistic units/ transmission of linguistic units	→	codeswitching/borrowing
analogical selection of linguistic units/ partial transmission of linguistic units	→	interference (transfer)
transmutation of linguistic units	→	conceptual and formal replicates

Firstly, when a speaker co-actively selects linguistic units from the different codes s/he has available, this can lead to codeswitching (including codemixing; cf. Muysken 2000). The term "borrowing" is often used to describe the process that linguistic units are completely transmitted to another code at the language-systemic level.[15] Secondly, particularly in contact settings where second (third, etc.) language acquisition takes place, "selection in the mental network can [...] occur in analogy to formally and/or conceptually closely related units across different codes" (Onysko 2016: 210). False friends are typical examples. Onysko (2016: 210) uses the term "interference" to account for the fact that analogical selection can occur below the

[15] Onysko (2016: 210) points out that after some debate, language contact theory has agreed on a continuum between codeswitching as a speaker-specific process and borrowing as the use of a contact feature in a speech community (e.g., Matras 2009: 110–114).

level of consciousness. "Transfer" is an alternative label when used in this strict sense (and not in reference to all kinds of contact influence). The last process the LCT distinguishes is the transmutation of linguistic units; thus, "a conceptual stimulus from code A is rendered into code B by using linguistic material from code B" (ibid.). Examples are replications in the sense of loan translations. Mixed forms of the abovementioned processes are possible.

Regarding the linguistic parameters underlying language contact, Onysko (2016: 211) points out that contact influence is measurable on the basis of "hierarchies of borrowability" (ibid.: 210). Nominal borrowings make the start and are followed by borrowings of grammatical or phonetic features. Consequently, the sole borrowing of a few nominal concepts is a sign of weak contact influence, whereas the borrowing of grammatical or phonetic features indicates relatively strong language contact. Socio-cultural extra-linguistic parameters Onysko (2016: 211) mentions are demographics, dominance, and discourse. Demographics comprises, among other aspects, the speaker number or the "areal spread" (ibid.) of the languages involved. Dominance refers to speaker status, the socio-economic power hierarchy of the contact languages, etc. Discourse describes the communicative situations and domains in which the contact is established. A second extra-linguistic parameter that influences the manifestation of contact features is speaker attitude towards those features. If the features do not correspond to the communicative intentions, idiolect, and the like of a speaker, s/he is unlikely to use them, but "parallel activation in the language network can lead to an unmonitored, spontaneous emergence of contact features, that is, transfer" (ibid.).

With the aim to "captur[e] general phenotypes that share certain characteristics" (ibid.: 212), Onysko (2016: 212–213) depicts the LCT of World Englishes as a circle that comprises five contact settings: global Englishes (GEs), learner Englishes (LEs), Englishes in multilingual constellations (EMCs), English-based Pidgins and Creoles (EPCs), and Koiné Englishes (KEs). This macro-level consideration of contact settings allows identification of prototypical settings that share certain characteristics.[16] GEs are examples of English as a global language in settings where English does not function as an official language or as an everyday means of communication (ibid.: 212). LEs are likely to co-occur with GEs in settings where the acquisition of English is part of the school curriculum but where English is not used in everyday

[16]The five contact settings are only broadly outlined here. For details on the dimensions that connect members of the respective settings, see Onysko (2016: 212–215).

communication. Speakers of LEs are likely to transfer features from their native language(s) to English, also when they use English as a lingua franca to communicate with speakers of a different mother tongue. In EMC settings, in contrast, English plays a significant role (for instance as an official language) and is either acquired as a first language or second (third, etc.) language. Englishes in postcolonial settings, where intense contact with local languages can result in mixed lects like Singlish, the local vernacular spoken in Singapore, are typical examples. EPCs are "codes with restricted functions" that have emerged in contact settings where Englishes "have lexified the contact language" (ibid.: 214). Additionally, non-dominant languages have contributed structural and lexical features. KEs, finally, emerge from dialect contact and comprise standard BrE and AmE as well as many British and American dialects. The typology takes account of historical dimensions, which Onysko (2016) considers "vital for a classification, as contact settings of varieties often change throughout their development" (214).

Onysko's (2016) LCT is attractive in that it incorporates multiple dimensions and considers both linguistic and extra-linguistic factors. In the introductory pages to her recently edited volume called *Modeling World Englishes*, Deshors (2018: 2) refers to Onysko (2016) by saying:

> [I]nvestigating the evolution of Englishes is a multifaceted phenomenon that requires researchers to account for not only linguistic factors but also historical and sociological ones. While those factors are themselves, constantly evolving with time, likewise, our theoretical models should also demonstrate ongoing developments by accounting for the ways that Englishes are used as a result of their global ongoing spread.

From a usage-based perspective, the inclusion of cognitive concepts such as analogy in Onysko's (2016) LCT is particularly noteworthy. It remains to be seen how the inclusion of cognitive processes underlying contact phenomena such as codeswitching or transfer in a multidimensional model is taken up by the World Englishes community and beyond. From a usage-based perspective it is certainly essential. The LCT will be applied to the varieties of interest in sections 3.2 (SgE), 3.3 (HKE), and 3.4 (IndE).

Bao's (2010) usage-based approach to substratum transfer

Bao's (2010) usage-based approach to substratum transfer, according to which "frequency of use in the contact language plays a crucial role in substratum-derived

linguistic change" (792), is worth considering here as well. According to Bao (2010: 792), New Englishes[17] are (just as pidgins and creoles) affected by grammatical change due to contact with local languages, but in contrast with pidgins and creoles "the continued presence of the local languages—the linguistic substratum—and English provides the necessary sociolinguistic condition for the type of contact-induced grammatical restructuring that is characteristic of New English varieties" (ibid.). Bao (2010: 793) argues that the contact literature has concentrated on identifying individual features that emerged in contact scenarios involving substratum transfer without paying attention to the productivity (i.e., the usage frequency) of those features in the contact language. Whenever frequencies of use have been considered, the focus has been on features that are frequent enough in the substratum language(s) to be transferred into the contact language (e.g., Siegel 1999; Mufwene 2001). Bao (2010) uses the term "substratum" to refer to "the language that contributes the grammatical features" (795) and the term "lexifier" to define "the language that provides the morphosyntactic exponence for the transferred features in the contact language" (795). Morphosyntactic exponence describes "the lexical, morphological, or syntactic materials used to express, or spell out, grammatical features transferred from the linguistic substratum" (ibid.). There is a tension in the stabilization process insofar as that the lexifier language needs to provide suitable morphosyntactic material to make a grammatical feature in the substratum language find its way into the contact language. I.e., the grammatical feature "must be filtered through the usage patterns that are ultimately determined by the lexifier language" (ibid.: 814).

Examples 3.1 and 3.2 (taken from Bao 2010: 814, example 38) depict the transfer process for the *give*-passive in SgE. Bao (2010: 801–804) argues that the SgE *give*-passive stems from the Hokkien *ho*-passive, *ho* ("give") being typically used in two morphosyntactic frames:

(3.1) *ho* NP_1 NP_2

(3.2) *ho* NP V

In 3.1, *ho* functions as a ditransitive verb, in 3.2 it is used in passive voice. On the basis of exemplar theory, which theorizes that tokens of a particular linguistic form are stored as exemplars of that linguistic form (e.g., Johnson 1996; Bybee 2001;

[17]Bao (2010) defines New Englishes as Englishes that "are typically found in historical settings with British colonial administration but without sizable English-speaking settlements" (792).

2002; 2006; Pierrehumbert 2001; Bybee & Eddington 2006; Bao 2009), Bao (2010: 814) assumes that tokens adopting these morphosyntactic frames are exemplars of these frames. The more frequent a particular exemplar is, the more readily available it is in the mental representation, i.e., the stronger it is. Each exemplar provides all linguistically relevant information (from the physical properties of the speech signal to the abstract representation of the frame; cf. Pierrehumbert 2001; Bybee 2006). While both *ho*-frames work in Chinese, only 3.1 works in English. English (the lexifier) provides the morphosyntactic material that allows for the transfer depicted in 3.1, but not the material that would allow for the transfer shown in 3.2. This is why frame 3.2 does not or hardly occur in SgE. The very low frequency of occurrence of the *give*-passive in morphosyntactic frame 3.2 will ultimately result in the absolute extinction of the use of the *give*-passive in that frame in SgE. Bao (2010) tests his theory by means of four unproductive grammatical features in SgE: the perfective cluster, the *kena* and *give* passives, serial verbs, and reduplication.

Bao (2010) describes Singapore as "a contact ecology, where all contributing languages are equally active and easily accessible" (ibid.: 812), the majority of Singaporeans being "English-knowing bilinguals" (a term coined by Pakir 1991), whose grammatical intuitions about the languages involved are likely to impact substratum transfer. It is an open question how far the mechanisms of Bao's (2010) usage-based account hold for speech communities in which English is not as equally active and as easily accessible as the local languages it is in contact with.

3.2 Singapore English

3.2.1 The institutionalization of Singapore English

Singapore is a tropical island state at the southern end of the Malay Peninsula comprising an area of about 700 km^2 (cf. Leimgruber 2013: 1). English reportedly came to Singapore in 1819, when Stamford Raffles annexed Singapore as a strategic trading post (cf. Bao 2015: 15). At that time, merely about 150 fishermen and pirates were living on the island which was to become a major economic player in the future. Under the administration of Calcutta, Singapore became a permanent British settlement (cf. Leimgruber 2013: 1), and immigration from the Malay Peninsula, Indonesia, and later from the southern Chinese provinces of Fujian and Guangdong started to flourish (cf. Bao 2015: 15).

Table 3.2 depicts the resident population of Singapore by ethnic group between 1840 and 2010, the Chinese population continuously being the biggest ethnic group. Since 1911, the ethnic distribution has been stable (cf. Bao 2015: 16). The nineteenth as well as the early twentieth century, which Bao (2015) calls "the formative period of Singapore English" (16), were characterized by constant population increase and movement, particularly in the Chinese community (cf. Hsu 1950, in Bao 2015: 17). Dramatic political changes in China resulting in the collapse of the Quing or Manchu dynasty in the early twentieth century and the opening of China towards Western markets made increasing numbers of Chinese immigrants settle permanently in Singapore. Schneider (2007) points out that already from 1867 onwards, when Singapore was annexed as a crown colony, the importance of its port as "an international trading center" (154) contributed to Singapore's growth. He defines this as the beginning of stage 2 ("endonormative stabilization") in his Dynamic Model. The colonial phase ended with the Japanese occupation of Singapore during World War II from 1942 to 1945, after which the British gained control over Singapore again. However, the spirit among the ruled had changed, which is a sign for stage 3 "nativization." This was particularly noticeable in the newly founded People's Action Party, which pushed for Singapore's independence.

Table 3.2: Resident population of Singapore by ethnic group from 1840 to 2010 (in percent; adopted from Bao 2015: 17)

year	population	Chinese	Malays	Indians	Others
1840	35,389	50.0	37.3	9.5	3.1
1860	81,734	61.2	19.8	15.9	3.1
1891	181,602	67.1	19.7	8.8	4.3
1911	303,321	72.4	13.8	9.2	4.7
1931	557,745	75.1	11.7	9.1	4.2
1957	1,445,929	75.4	13.6	8.6	2.4
1980	2,413,945	76.9	14.6	6.4	2.1
1990	2,705,115	77.7	14.1	7.1	1.1
2000	3,273,363	76.8	13.9	7.9	1.4
2010	3,771,721	74.1	13.4	9.2	3.3

Sources: Pan (1999), Saw (1999), Department of Statistics Singapore (1980; 1990; 2000; 2010); all in Bao (2015)

With the different ethnic groups settling down in Singapore, a range of dialects from diverse language families came to the island. The Chinese immigrants spoke different Chinese dialects that are partly mutually unintelligible on mainly phono-

logical grounds, namely Hokkien, Teochew, Cantonese, Hakka, and Hainanese (cf. Bao 2015: 18–19). Mandarin, in contrast, was not among the dialects the early immigrants brought along. The Peranakans (also part of the Chinese community) who came from the Straits Settlements, spoke Baba Malay, a creole based on Malay with substratum influence from Hokkien. Of the Indian immigrants, 93 percent spoke Dravidian languages, namely Tamil, Malayalam, and Telugu, and seven percent Indo-European languages, namely Hindi, Punjabi, and Bengali (cf. Walker 2004, in Bao 2015: 18). The lingua franca outside the geographically separate ethnic enclaves in early Singapore was Bazaar Malay, a Malay-based pidgin, but a form of pidginized English overtook Bazaar Malay in the 1970s (cf. Bao 2015: 21).

Education significantly contributed to the spread of English, although Bao (2015) points out that it "played a limited role in the emergence and stabilization of the grammar of Singapore English, especially the part of its grammar that demonstrably derives from the local languages" (25). Compared with private conversations where the vernacular dominates, school has rather been the environment to encounter "the formal, scholastic variety" (ibid.). English-medium education came with Christian missionaries and was only available to a small elite (cf. Leimgruber 2013: 3), but enrollment into English-medium schools surpassed enrollment into Chinese-medium schools after World War II. After three years of belonging to the Federation of Malaya (1962 to 1965), Singapore gained independence in 1965. The post-independence government, with Lee Kuan Yew's People's Action Party occupying the majority of the Parliament's seats (cf. Turnbull 1977: 263), pushed an "English-centric bilingual education policy" (English plus mother tongue; cf. Bao 2015: 27), and promoted Mandarin as the shared dialect among ethnically Chinese Singaporeans. The *Speak Mandarin Campaign* was launched in 1979, which is likely to have contributed to decreasing uses of Chinese dialects as the main home language (compare table 3.3 below). Since 1987, all school subjects have been taught in English and the mother tongues (Mandarin, Malay, and Tamil) have been treated as a second or foreign language (cf. Bao 2015: 29), depending on whether the home language is one of the official mother tongues or not. According to Schneider (2007: 155), the vast spread of English from 1965 (the year of independence) onwards marks the transition into stage 4 ("endonormative stabilization").

The constitution defines Malay, Mandarin, Tamil, and English (in that order) as "the four official languages of Singapore" (Constitution 1965: §153A, in Leimgruber 2013: 7). Malay is additionally the "national language" (§153A), meaning that the

national anthem and drill commands are in Malay. In the ethnically Chinese group, Mandarin is being increasingly used when talking to fellow Chinese Singaporeans, with the side effect that younger Chinese Singaporeans face difficulties when they are supposed to talk to their grandparents in dialect (cf. Leimgruber 2013: 8). The Indian Singaporean speech community comprises Tamils, Malayalees, Telugus (speakers of Dravidian languages) as well as Bengalis, Punjabis, and Sinhalese (speakers of Indo-Aryan languages) and has lost nothing of the linguistic diversity of the first Indian immigrants coming to Singapore (see above). It faces the challenge that Tamil is not only just one of the Indian languages spoken in Singapore, but it is also of limited importance in daily life apart from having been the home language of a mere 37.7 percent of the Indian Singaporean population in 2015 (compare table 3.3 below) and being displayed on signs or used in (public transport) announcements wherever Singapore's language policy prescribes all four official languages.

Table 3.3 shows the resident population of Singapore by ethnic group and language most frequently spoken at home for the years 2000, 2010, and 2015. The data for 2000 and 2010 are taken from the *Census of Population 2010* (Department of Statistics Singapore 2010), those for 2015 from the *General Household Survey* (Department of Statistics Singapore 2015). The latter provides no data for the speaker group "Others" in 2015. The resident population comprises both Singapore citizens and permanent residents. Leaving the group "Others" aside, use of English as the main home language is highest in the Indian population, followed by the Chinese and Malay populations, with English being increasingly used in all three groups. The Malays show the most homogeneous language behavior by mainly using Malay at home. The Indians use Tamil to a similar extent as English, and Malay and other languages to a considerable degree. In the Chinese group, more people report Mandarin than English to be their main home language, followed by Chinese dialects other than Mandarin, whose use at home is declining.

An additional look at the distribution of main home languages across ethnic groups makes sense because Singapore's ethnic groups differ considerably in size (compare table 3.2). Across ethnicities ("total"), it is the use of Chinese dialects that has particularly declined between 2000 and 2015. In contrast, uses of Mandarin, Malay, and Tamil have remained stable. The use of Chinese dialects declined between 2000 and 2010 from 23.8 percent to a mere 14.3 percent, whereas English was used as the main home language by 23.0 percent in 2000 compared with 32.3 percent in 2010. Mandarin (2000: 35.0 percent, 2010: 35.6 percent), Malay (2000: 14.1 percent,

Table 3.3: Resident population of Singapore aged five years and over by ethnic group and language most frequently spoken at home (percentages; Department of Statistics Singapore 2010: Key Indicators: Literacy and Language; Department of Statistics Singapore 2015: 19, Chart 3.3)

year	English	Mandarin	Chin. dialects	Malay	Tamil	Others	sum
all:							
2000	23.0	35.0	23.8	14.1	3.2	0.9	100.0
2010	32.3	35.6	14.3	12.2	3.3	2.3	100.0
2015	36.9	34.9	12.2	10.7	3.3	2.0	100.0
Chinese:							
2000	23.9	45.1	30.7	0.2	—†	0.1	100.0
2010	32.6	47.7	19.2	0.2	—†	0.2	100.0
2015	37.4	46.1	16.1	‡	‡	0.4	100.0
Malays:							
2000	7.9	0.1	0.1	91.6	0.1	0.3	100.0
2010	17.0	0.1	—†	82.7	0.1	0.2	100.0
2015	21.5	‡	‡	78.4	‡	0.1	100.0
Indians:							
2000	35.6	0.1	0.1	11.6	42.9	9.7	100.0
2010	41.6	0.1	—†	7.9	36.7	13.6	100.0
2015	44.3	‡	‡	5.6	37.7	12.4	100.0
Others:							
2000	68.5	4.4	3.2	15.6	0.2	8.2	100.0
2010	62.4	3.8	0.9	4.3	0.1	28.6	100.0

Note: The order of ethnic groups and main home languages follows that in the census, and the *General Household Survey* does not provide data for Others in 2015.

† The sources provide no data.

‡ The percentages are included in "Others."

2010: 12.2 percent), and Tamil (2000: 3.2 percent, 2010: 3.3 percent) have remained stable. Thus, it seems that the use of English as the main home language has gained ground at the expense of the use of Chinese dialects in particular.[18] As Leimgruber (2013: 8) notes, the fact that only 20.2 percent of all marriages were inter-ethnic marriages in 2010 shows that inter-ethnic marriages alone are unlikely to explain the increasing use of English at home.

According to Schneider (2007), "Singapore English has gone through a vibrant process of structural nativization, more visibly on the basilectal level but also in formal styles" (158), which is why he clearly assigns Singapore stage 4 in his Dynamic Model (ibid.: 160). A clear national multiethnic identity prevails that expresses itself in a conscious use of the local vernacular Singlish in situations that allow for it. This vernacular combines phonetic, grammatical, and lexical features that make Singlish unique and creates a feeling of belonging and national pride among Singaporeans (cf. Schneider 2007: 160). Ooi (2001a) even sees signs of stage 5 ("differentiation") in the emergence of distinct ethnic varieties of SgE. Use of Singlish is met by continued governmental attempts to limit its spread and promote the use of "good" (i.e., standard British) English (cf. Bao 2015: 35–36). The government even launched a campaign in 2000 called the "Speak Good English Movement." It has been repeatedly shown, however, that Singaporeans perceive Singlish as a variety whose use is constrained to informal settings like chatting with family or friends (e.g., Chng 2003).

SgE is one of the Englishes in multilingual constellations (EMCs) in Onysko's (2016) LCT of World Englishes. English plays an important role in all domains of life and functions as an inter-ethnic lingua franca among the different ethnicities living in Singapore. The intense contact between English and the local languages manifests itself in the local vernacular and identity carrier Singlish (Onysko 2016), and Bao (2015) explains various variety-specific features with systemic transfer from Chinese substrata. Increasing use of English as the main home language has come at the expense of Chinese dialects other than Mandarin in particular (table 3.3), with unpredictable consequences for the linguistic diversity and language contact situation in Singapore in the long run. Nevertheless, SgE is so widely used in everyday life that contact features continue to develop and stabilize not only among single speakers but in the entire speech community (cf. Onysko 2016: 206).

[18] As mentioned above, today use of dialects is often restricted to communication among or with the elderly (cf. Leimgruber 2013: 8).

As pointed out before, Bao (2015: 25) emphasizes that English-medium education was of minor importance for the development and stabilization of grammatical features of SgE that derived from local languages. "[P]eople [have] acquire[d] the informal, vernacular variety at home or on the street" (ibid.: 25) instead. In line with that, Onysko (2016) calls the attitude speakers have "towards language behavior [...] the decisive criterion that regulates the occurrence of language contact" (211) and the use of contact features as "a contextually-bound reaction that reflects the speaker's communicative intentions" (211). I.e., it is the speakers' active choice that determines the development and stabilization of the local vernacular Singlish.

3.2.2 Substrate influence on Singapore English

Bao (2015) considers the years from the mid-nineteenth to the mid-twentieth century "the formative period of Singapore English" (33) because English developed towards the lingua franca of Singapore then. He lists the following languages as "major heritage languages of Singapore": English, Chinese dialects other than Mandarin (Hokkien, Teochew, Cantonese, Hainanese, Hakka, Foochow, other dialects), Mandarin, Malay, Tamil and other languages of India, and the contact languages Bazaar Malay (Malay-lexified pidgin) and Baba Malay (Malay-lexified patois of the Paranakans). While the Chinese population comprised few native speakers of Mandarin in the 1950s, "[the] place [of Mandarin] in the contact ecology of twentieth-century Singapore is beyond doubt" (Bao 2015: 34). In the *Singapore Census of Population* 1957, the native speakers of Mandarin are part of the category "all other dialects," which 2.6 percent of the population spoke. Bao (2015: 34) describes the language situation in the Chinese community as diglossic: Mandarin functions as the "high variety" and the "Chinese dialects" as "low varieties," which is a result of Singapore's mother-tongue-based education policy. In contrast with the complementary functions "scholastic English" and SgE perform, Mandarin has increasingly taken over the traditional role of the major dialects Hokkien, Teochew, and Cantonese as a home language (Bao 2015: 34). Lately, English has gained importance as a home language among members of the Chinese community.

Bao (2015) stresses the "clear and irrefutable Chinese influence on the grammar of Singapore English in particular" (35). He reasonably argues that the Chinese do not only comprise the by far largest ethnic group, but their use of English as the main home language has steeply increased since the 1980s and more than in the other ethnic groups. In the *Singapore Census of Population* 1980 (Department of Statistics

Singapore 1980), 7.6 percent of the Chinese resident population aged five years and older reported that they mainly use English at home compared with 19.3 percent in 1990, 23.0 percent in 2000, and 32.6 percent in 2010 (Department of Statistics Singapore 1980; 1990; 2000; 2010).[19] In that context, it is worth pointing out that "[d]espite the varying degrees of intelligibility among the dialects, it is generally accepted within the Chinese linguistics circle that the dialects share a common core in grammar and vocabulary (Chao 1968)" (Bao 2010: 794). Thus, if substratum transfer promotes the omission and regularization phenomena considered in this book, common tendencies can be expected from the Chinese dialects represented in Singapore. The contact languages Bazaar Malay and Baba Malay have had negligible influence on SgE according to Bao (2015: 35). Bao (2015: 35–36) raises the important issue that the local vernacular is stigmatized in various respects and points out that stigmatized features spread relatively slowly (e.g., Labov 1972). He puts it as follows: "[T]he grammatical features which reflect the influence of Chinese or other local languages bear the brunt of the stigma" (36). Examples are stigmatized sound changes in SgE that do not affect all possible candidates, which is why they are little spread and instable.

3.3 Hong Kong English

3.3.1 The institutionalization of Hong Kong English

Bolton (2002b: 31) traces the origins of HKE back to the arrival of the first British trading ships at neighboring Macau and Canton (Guangzhou) in the early seventeenth century. Trading in the region led to the emergence of "a distinct variety of Chinese pidgin English" (ibid.) spoken in Canton and Macau, which was referred to as "jargon" in the beginning and as "pidgin English" from the 1860s onwards. In 1842, during the First Opium War between Britain and China, Hong Kong was annexed as a British colony, and the establishment of missionary schools in Hong Kong and China led to the spread of English among small parts of the indigenous population (cf. Bolton 2002b: 31–32; Schneider 2007: 135).[20] At the same time, the

[19] Among the Malay residents, 1.5 percent indicated that they use English as the main home language in 1980, 6.1 percent in 1990, 7.9 percent in 2000, and 17.0 percent in 2010. In the Indian group, the respective numbers were 18.9 percent in 1980, 32.3 percent in 1990, 35.6 percent in 2000, and 41.6 percent in 2010.

[20] At that time, Hong Kong only had approximately 7,000 inhabitants that mainly lived in small fishing villages on the southern shore (cf. Evans 2014: 580). Up until 1941, Hong Kong contin-

study of Chinese remained important because of both the "strong literary and philosophical tradition" (Bolton 2002b: 32) in the region and the missionaries' interest in Mandarin, Cantonese, Hakka, and Chiu Chau. The non-Chinese members of society, among them the English-speaking "settlers", only made up two to five percent of Hong Kong's population then (cf. Evans 2014: 579). Hong Kong is a special case compared with other former colonies in that it lacks linguistic diversity, the reason being that Cantonese has been the main language spoken for a considerable amount of time (cf. Bolton 2003). English has never served as a lingua franca in Hong Kong, a communicative function it typically has among members of the indigenous strand according to Schneider's Dynamic Model (cf. Schneider 2007: 35–36, 67). Schneider (2007: 135) sees evidence of the transition to stage 2 ("exonormative stabilization") in the Treaty of 1898, which leased the New Territories for ninety-nine years and guaranteed British hold of the area. Bilingualism spread among elitist parts of society that acquired English at school, and the British expatriates saw themselves as "representatives of Britain in an Asian outpost" (ibid.: 135), while the Hong Kong people increasingly came in contact with the culture the expatriates brought along (ibid.; but see the critical objection by Evans 2014: 581 below).

The University of Hong Kong was established in 1911 as an English-medium university. Nevertheless, many Chinese-medium schools were founded in the 1920s and 1930s (cf. So 1992: 72) and the Chinese University of Hong Kong was set up in 1963 (cf. Bolton 2002b: 32), which indicates that Chinese maintained its importance in the education sector in the twentieth century. The "Chinese language campaign" in the 1970s claimed stronger recognition of Chinese and was followed by education reforms in the mid to late 1970s at the primary and secondary level. It is worth mentioning in that context that Hong Kong witnessed a large population increase after the establishment of the People's Republic of China in 1949 (600,000 inhabitants in 1945 compared with 3.1 million in 1961; cf. Bolton 2002b: 32). According to Evans (2014) "these refugees became 'settlers' (1950s–1960s) [in Schneider's terms] and their children 'Hongkongers' (1970s–1980s)" (581), and only from the 1960s onwards the (slight) majority of Hong Kong's Chinese population has actually been born in Hong Kong (582). The education reforms of the 1970s led to the diffusion of English because it allowed a large proportion of children in "Anglo-Chinese" secondary schools to acquire English (cf. Bolton 2002b: 34; Evans 2014: 592).

ued to be mainly populated by Chinese transients on the one hand and fishing and farming communities on the other hand.

Schneider (2007: 135) marks the beginning of stage 3 ("nativization") with "the economic transformation of Hong Kong [in the 1960s] from a relatively poor refugee community to a wealthy commercial and entrepreneurial powerhouse" (Bolton 2000: 268, in Schneider 2007: 135). The Joint Sino-British Declaration of 1984 resulted in the handover of Hong Kong to the People's Republic of China in 1997 for a transition period of fifty years. For the identity constructions of members of the settler strand this meant that they were now "permanent Hong Kong resident[s] of British origin" (Schneider 2007). Further evidence of the transition to stage 3 is that "[f]or at least thirty years, Hong Kong has had its own localized complaint tradition about 'falling standards' of both English and Chinese" (Bolton & S. G.-l. Lim 2002: 298, in Schneider 2007: 137).

The variety status of HKE has been a matter of debate in the (recent) past (Schneider 2007: 137). In the 1980s, Luke and Richards (1982: 55) assigned English in Hong Kong an exonormative orientation and spoke out against the existence of a distinct "Hong Kong English," an attitude that is still found today among teachers in Hong Kong (cf. Tsui & Bunton 2000). Bolton & Kwok (1990) describe segmental and supra-segmental features of HKE phonology, showing that speakers of HKE in fact "share a number of localised features of a Hong Kong accent" (166). Bolton (2002b: 49) points out that he does not wish to promote the existence of a distinct variety though. Evans (2014) remains skeptical of the development of "a localised variety of English" because "the use of English is rather limited in Cantonese-speaking Hong Kong, and thus there is no societal basis for the development of a nativised variety" (ibid.: 592; compare also Evans 2011). He cites Joseph (2004), according to whom "international recognition [of variety-specific features] has come in the almost total absence of local assertion" (Evans 2011: 148). Schneider (2007), in contrast, provides a number of distinct vocabulary items and lexicogrammatical features of HKE. Examples of the former are loans or interferences from Cantonese or other Chinese dialects such as *kwailo* "foreign residents" or *char siew* "Chinese-style roast pork" and loan translations like *blue lantern*; an example of the latter is the use of uncountable nouns as countable nouns (e.g., *equipments*, *aircrafts*).

Language planning in the Hong Kong Special Administrative Region has evolved around uses of Cantonese, Putonghua, and English. While Chinese was recognized as a co-official language in 1974, English served as the *de facto* official language of government and law and was mainly used in education as well as in trade, business, and finance during British colonial rule. With the "Handover" of Hong Kong to the

People's Republic of China (PRC) in 1997, the role of Chinese further strengthened. This boosted conflicts between uses of Cantonese and Putonghua. The language policy of the PRC has officially discouraged use of dialects such as Cantonese, while the promotion and use of Cantonese is a legacy of colonial language policy (Bolton 2002b: 37).

Shortly before the "Handover" in 1997, English-medium schools got suddenly and strictly limited in number and most secondary schools had to stick with or switch to Cantonese as the medium of instruction (ibid.: 38–39). Since then, reassurances by the government concerning the study of English have not stopped parents from worrying that the relegation of the status of English to that of a foreign language in the curriculum is a disadvantage for their children, with good English skills becoming ever more important for future (job) prospects (ibid.: 40).

Let us consider recent census data in order to get a better idea of current patterns of language use in Hong Kong. The *2016 Hong Kong Population By-census* provides information on the languages or dialects members of the Hong Kong population aged five and above usually and additionally use. The results are summarized in table 3.4. The By-census defines "the language/dialect a person used in daily communication at home" as the "usual language" (Census and Statistics Department Hong Kong 2017: 139).

Table 3.4 shows that in 2006, 2011, and 2016 the large majority of the population used Cantonese as the usual language for daily communication at home. Both English and Putonghua were mainly used as another language or dialect, respectively, with English outperforming Putonghua in 2006 and 2016 and Putonghua outperforming English in 2011. Note the considerable rise in use of English as another language or dialect in 2016. Other Chinese dialects and other languages (Bahasa Indonesia, Filipino, and Japanese) were hardly used as either the usual language or another language or dialect. The *Thematic Household Survey Report Nr. 59* (Census and Statistics Department Hong Kong 2016) is worth mentioning in that context. Employed persons aged 15 to 65 were asked whether they would learn or further study Cantonese, spoken English, or Putonghua, as well as written Chinese or written English for the sake of work. 73.3 percent responded that they would learn or further study spoken English compared with 17.4 percent that would learn or further study Putonghua; 9.3 percent mentioned Cantonese, 91.8 percent written English, and 8.2 percent written Chinese. These percentages show that people in Hong Kong

Table 3.4: Proportion of population aged five and over able to speak selected languages/dialects (adopted from the Census and Statistics Department Hong Kong 2017: 46)

language/dialect	as the usual lan.			as another lan./dialect			total		
	2006	2011	2016	2006	2011	2016	2006	2011	2016
Cantonese	90.8	89.5	88.9	5.7	6.3	5.7	96.5	95.8	94.6
English	2.8	3.5	4.3	41.9	42.6	48.9	44.7	46.1	53.2
Putonghua	0.9	1.4	1.9	39.2	46.5	46.7	40.2	47.8	48.6
Hakka	1.1	0.9	0.6	3.6	3.8	3.5	4.7	4.7	4.2
Fukien	1.2	1.1	1.0	2.1	2.3	2.6	3.4	3.5	3.6
Chiu Chau	0.8	0.7	0.5	3.2	3.1	2.9	3.9	3.8	3.4
Indonesian (Bahasa Indon.)	0.1	0.3	0.3	1.5	2.2	2.4	1.7	2.4	2.7
Filipino (Tagalog)	0.1	0.2	0.4	1.3	1.4	2.3	1.4	1.7	2.7
Japanese	0.2	0.2	0.1	1.1	1.4	1.7	1.2	1.5	1.8
Shanghainese	0.3	0.3	0.2	0.9	0.9	0.9	1.2	1.1	1.1

Note: Mute persons are excluded. The total numbers for the usual languages are provided (and partly accumulated) in the *2016 Hong Kong Population By-census*.

are well aware of the importance of English and Putonghua for work, and that they are willing to improve their respective language skills.

According to Joseph (2004: 159–161, in Schneider 2007: 139), HKE might gain importance as an identity carrier, particularly in case the Beijing government should increasingly try to make Hong Kong people adopt a northern Chinese identity by pushing the role of Putonghua in Hong Kong. It remains to be seen whether Putonghua and English will gain strength in the future. Results of the *Thematic Household Survey Report Nr. 59* certainly show that people in Hong Kong are well aware of the roles English and Putonghua play globally (see above). While HKE has a number of typical features, it lacks characteristics that make the variety unique (compared with SgE, for instance). As long as HKE does not stabilize (further) or functions as an identity carrier, a move to stage 4 of the Dynamic Model ("endonormative stabilization") is unlikely.

In Onysko's (2016) LCT, HKE clearly constitutes one of the Englishes in multilingual constellations (EMCs). Although English is outperformed by Cantonese as the means of daily communication, its official status and its use as "another language" besides Cantonese (table 3.4) make HKE an EMC. Typical "processes" in which En-

glish is acquired as an L2 are analogical selection, particularly when the languages in contact are typologically similar, and the partial transmission of linguistic units (e.g., in the form of systemic substratum transfer; cf. Bao 2015; Onysko 2016: 210). Due to the colonial history of Hong Kong, contact between English and the local languages has been long and intense, particularly from the education reforms of the 1970s onwards, which provided large proportions of children immediate access to English (Bolton 2002b: 34; Evans 2014: 592). Nevertheless, the fact that Cantonese largely prevails in most matters of daily life prevents contact features from flourishing more quickly or intensely; both within single speakers and in the entire speech community (cf. Onysko 2016: 206).

3.3.2 Substrate influence on Hong Kong English

As the census data show, Cantonese is by far the most frequently used dialect in Hong Kong, followed by English and Putonghua (compare table 3.4). Historically, the large numbers of transients from neighboring regions outnumbered the non-Chinese (among them the English-speaking "settlers") by far at any point in time (cf. Evans 2014: 580), and population figures rose steeply after 1949, when numerous refugees from China settled in Hong Kong after the establishment of the People's Republic of China. Putonghua (and its conflicts with Cantonese) gained importance with the handover of Hong Kong to the People's Republic of China in 1994 only, but the widespread use of Cantonese in all matters of daily life has made it difficult for Putonghua to gain ground and be officially promoted as the dialect to aim for. While English-medium education became available to the masses after the education reforms of the 1970s, Cantonese has regained its role as the main medium of instruction after the handover of Hong Kong to the People's Republic of China.

Research on substrate influence on HKE is limited to feature-based accounts[21], but the fact that "Cantonese has long been the majority language (Bolton 2003)" (Evans 2014: 579) leaves little doubt that it is the main substratum of HKE. As pointed out in the description of SgE above, Chinese dialects "share a common core in grammar and vocabulary" (Bao 2010: 794) despite the fact that they are partly mutually unintelligible. This means that influence besides Cantonese from Mandarin or other Chinese dialects spoken in Hong Kong, such as Hakka, Fukien, or Chiu Chau (table 3.4), is unlikely to change the picture drastically.

[21]Compare, for instance, Gisborne's (2009) study on the morphosyntactic typology of HKE.

3.4 Indian English

3.4.1 The institutionalization of Indian English

India is the largest country in the Indian subcontinent. It borders (from west to east) Pakistan, China, Sri Lanka, Nepal, Bangladesh, Bhutan, and Myanmar (formerly Burma) and ranges over an area of 3.3 million square kilometers (cf. Sailaja 2009: 1). India is "administered through a loosely federal form of government (ibid.) that regulates 28 states and seven mostly linguistically determined union territories. Its linguistic diversity is manifold, with languages spoken in India belonging to four different language families: Indo-Aryan (e.g., Hindi, Bengali, Urdu), Dravidian (e.g., Tamil, Telugu, Malayalam), Tibeto-Burman (e.g., Angami, Bodo), and Austro-Asiatic (e.g., Munda, Khasi). The Indo-Aryan languages are spoken in the northern and eastern parts of India, the Dravidian languages in the southern parts, and the Austro-Asiatic and Tibeto-Burman languages in the eastern parts.

English first came to India more than four hundred years ago as the language of early missionaries, settlers, and merchants (cf. Sedlatschek 2009: 1). From the seventeenth century onwards, the East India Company began trading with India on the basis of a charter it had received from Queen Elizabeth. English served as the means of communication among traders that came to the region, so access to English was restricted to few Indians in the beginning (cf. Sedlatschek 2009: 9), and English was likely considered a foreign language (cf. Mukherjee 2007). The establishment of several missionary schools in the seventeenth and eighteenth centuries promoted the spread of English though (cf. Schneider 2007: 163). Britain gained control over the region after the Battle of Plassey in 1757, which the East India Company won against the last (independent) provincial viceroy of Bengal (cf. Schneider 2007: 163).

With the India Act of 1784, Britain brought the East India Company under full political control (cf. Sedlatschek 2009: 10). At that time, English language teaching started to rapidly spread in India, which peaked in Macaulay's *Minute on Indian Education* (1835) that successfully claimed English-medium education in India with the purpose to bridge communication gaps between the British colonizers and the indigenous population by educating parts of the population in English (cf. Schneider 2007: 163). Mukherjee (2007) considers Macauley's *Minute on Indian Education* as "the first step toward the beginning of the nativization of the English language in India" (165). Besides education, domains such as administration and the print media increasingly adopted English and have been run in English since then (Sedlatschek

2009: 11). Linguistically, the period saw increasing lexical borrowing from various Indian languages into English (such as *bamboo* or *dhoti* "loin-cloth worn by men", the former having found its way into international English, the latter remaining an example of local use; cf. Schneider 2007: 165). Schneider (2007) sets the start of stage 3 ("nativization") to the beginning of the twentieth century because "both the fundamental rooting of the language in the country and the emergence of its structural peculiarities must have predated the year 1947" (166), when India gained independence. One last major voice against the role of English in India was Gandhi, who stood up for the use of Hindustani rather than English as the "language of unity" (Schneider 2007: 166) across India.

With the declaration of independence in 1947, English became recognized as one of the official languages of India alongside Hindi for a transition period until 1965. However, the by then deep rooting of English in administration, education, and legislation made it difficult for Hindi to take over the role of English. Additionally, speakers of Dravidian languages in southern India in particular increasingly resisted the prospect of Hindi being the sole national language, arguing in favor of English as the "more neutral linguistic choice" (Sedlatschek 2009: 18). In 1963, the Official Languages Act paved the way for continued use of English as one of the official languages in India (Vaish 2008: 19, in Sedlatschek 2009: 19). The so-called "three language formula," established in 1968, which required education in Hindi (for non-Hindi speakers), in another modern Indian language (for speakers of Hindi), in English, and in one major regional language (the mother tongue), led to further protests among speakers of the Dravidian languages against learning Hindi (cf. Schneider 2007: 166; Sedlatschek 2009: 20). Similarly, speakers of Hindi spoke out against learning one of the Dravidian languages. This clearly shows the difficulties of implementing a nation-wide language education scheme in a linguistically diverse setting.

Today, English is no longer a compulsory but an alternative medium of education in primary and secondary schools, and it enjoys high approval in the population for the future (job) prospects it offers (cf. Sedlatschek 2009: 20). The local form of English does not function as an identity-carrier in India, but "it serves its classic role in an ESL country, that of an interethnically neutral link language which is qualified as a public and semiofficial language precisely because it is nobody else's (or at least not the competing group's) mother tongue" (cf. Schneider 2007: 167).

Obviously, India is not comparable in size with Hong Kong or Singapore, which likely has had a significant impact on the development of IndE as a distinct variety,

the development of regionally specific vernaculars, and identity-building. According to Schneider (2007), "an intersection of some non-regional features of educated pronunciation, [sic] perhaps represents an approximation to a uniform national standard" (172); a view also supported by Gargesh (2004: 992). Schneider (2007) points out, however, that "[a]t present it seems unlikely [...] that the language is going to cross the line and acquire new, emotionally more laden functions in Indian society" (173). Mukherjee (2007) interprets the ongoing protests against Hindi in southern India as an "Event X" in Schneider's terms (2007), i.e., as an incident that makes members of the settler strand "reconsider and redefine their position and [...] reconstruct a radically new, locally based identity for themselves" (Schneider 2007: 49). According to Mukherjee (2007), "the language riots made the political parties readjust their stance on language policy and ensure the continuing use of the English language in India" (168). He argues that English in India has proceeded to stage 4 ("endonormative stabilization") in Schneider's Dynamic Model in that "the process of nativization [...] is more or less over" (170). Mukherjee (2007) speaks of a variety that "is now largely endonormatively stabilized" (170) and "still relatively homogeneous" (ibid.), which is why differentiation (stage 5) has not yet taken place.

India counts 22 scheduled languages that are listed in the Eighth Schedule to the Constitution of India. Those scheduled languages are further subdivided into so-called "mother tongues." The "mother tongues" add up to 234 languages and are supplemented by 100 non-scheduled languages. Of the 22 scheduled and 100 non-scheduled languages, 24 are Indo-European languages (21 Indo-Aryan, two Iranian, one Germanic), 17 Dravidian, 14 Austro-Asiatic, 66 Tibeto-Burmese, and one Semito-Hamitic. Table 3.5 lists the 22 scheduled languages, the language family they belong to, the number of speakers who provided them as their mother tongue in the *2011 Census of India*, and the percentage of mother tongue speakers in India's population.[22] Hindi is the language with the by far highest number of mother tongue speakers, followed by Bengali and Marathi. Apart from Hindi, the scheduled languages listed are the mother tongue of very small percentages of the population, which is a sign of the enormous linguistic variation in India. The majority of the languages listed in table 3.5 are Indo-Aryan languages (I-A), followed by Dravidian languages (D), two Tibeto-Burmese (T-B), and one Austro-Asiatic language (A-A).

[22]There is a mistake with the decimal separators in the 2011 census data. For instance, the number of mother tongue speakers of Hindi is depicted as 52,83,47,193 instead of 528,347,193. In comparison, the *2001 Census of India* mentioned 422,048,642 Hindi mother tongue speakers.

Table 3.5: Scheduled languages in descending order of mother tongue (MT) speakers' strength (adopted from Office of the Registrar General and Census Commissioner, India 2018a and adjusted)

language	family†	MT speakers	%‡	language	fam.†	MT speak.	%‡
Hindi	I-A	528,347,193	43.63	Assamese	I-A	15,311,351	1.26
Bengali	I-A	97,237,669	8.03	Maithili	I-A	13,583,464	1.12
Marathi	I-A	83,026,680	6.86	Santali	A-A	7,368,192	0.61
Telugu	D	81,127,740	6.70	Kashmiri	I-A	6,797,587	0.56
Tamil	D	69,026,881	5.70	Nepali	I-A	2,926,168	0.24
Gujarati	I-A	55,492,554	4.58	Sindhi	I-A	2,772,264	0.23
Urdu	I-A	50,772,631	4.19	Dogri	I-A	2,596,767	0.21
Kannada	D	43,706,512	3.61	Konkani	I-A	2,256,502	0.19
Odia	I-A	37,521,324	3.10	Manipuri	T-B	1,761,079	0.15
Malayalam	D	34,838,819	2.88	Bodo	T-B	1,482,929	0.12
Punjabi	I-A	33,124,726	2.74	Sanskrit	I-A	24,821	0.00

†I-A: Indo-Aryan, D: Dravidian, A-A: Austro-Asiatic, T-B: Tibeto-Burmese (Government of India 2016)
‡Population size in 2011: 1,210,854,977 people (Office of the Registrar General and Census Commissioner, India 2018b: Statement 5)

As SgE and HKE, IndE is one of the Englishes in multilingual constellations (EMCs) in Onysko's (2016) LCT of World Englishes, and India's size and linguistic diversity make the country a promising test case for the contact typology. In a population that speaks various typologically different mother tongues (table 3.5), English does not only come in contact with many languages but also serves as a typical lingua franca (cf. Sedlatschek 2009: 20).

3.4.2 Substrate influence on Indian English

The enormous linguistic diversity in India makes the identification of substrate influence on IndE a challenging endeavor. Certainly, Hindi is a prime candidate for substratum transfer to English because of its high number of speakers (compare the number of people who indicated in the *2011 Census of India* that Hindi is their mother tongue; table 3.5). Scholars working on IndE find evidence of substratum transfer from various Indian languages or language families. To mention just a few, Carls (1999: 150), for instance, observes the word-formation pattern N-*cum*-N in IndE, which he relates to a Hindi pattern consisting of word stems that are coordi-

nated by the element *aur* ("and"). S. N. Sridhar (1992: 142–143) argues that IndE uses of *call/rename/term as* can be traced back to Dravidian languages, which mark names, technical terms, or quotations by the quotative particle (cf. Sedlatschek 2009: 166–174). Sedlatschek (2009: 202–227) investigates article usage in student essays in the *Primary Corpus*[23] and observes uses of the definite article where BrE or AmE has none (e.g., before proper names), but the use of zero article in places native Englishes require them prevails (cf. ibid.: 203). This speaks against substrate influence from Hindi on the article system of IndE. Hindi marks the definiteness of a noun not by means of a definite article but by word order (cf. Sharma 2005: 537).

[23]The *Primary Corpus* consists of 180,000 words of both spoken and written IndE and was compiled by Sedlatschek in 2000 (cf. Sedlatschek 2009: 41). It comprises press texts (40 texts), published broadcast material (40 texts), and student essays (ten texts) of 2,000 words each.

4 Data and methods

This chapter provides an overview of the data and methods used in the corpus analyses presented in chapters 5, 6, and 7, and in the experiment elaborated on in chapter 8. Detailed information precedes the analyses in the respective chapters. While the corpus analyses provide empirical accounts of the omission and regularization features of interest from a production perspective, the experiment investigates the perception of omission of inflectional past tense marking and omission of inflectional noun plural marking. The use of web-based corpus data to measure production and that of a web-based experiment to measure perception receive particular attention.

4.1 Corpora and corpus analyses

The corpus studies deal with omission of verbal past tense marking (chapter 5), omission of nominal plural marking (chapter 6), as well as regularization of irregular past tense marking in verbs and uses of uncountable nouns as countable nouns (both chapter 7) in the three Asian varieties of English of interest. The historical input variety BrE serves as a control. From a language contact perspective and given the relatively recent (social) media-related exposure of speakers of English around the world to AmE, it makes sense to account for AmE as a second control variety. However, the ICE corpora, which the corpus studies on omission of verbal past tense and nominal plural marking are based on, contain data from the 1990s and (early) 2000s, when the spread of AmE was not as far reaching as it is today.[24] This is why AmE is no valid control. In contrast, the corpus studies on regularization of irregular past tense marking in verbs and uses of uncountable nouns as countable nouns are based on GloWbE, which contains recent web-based data from 2012 (see below). GloWbE US was used as a control corpus in the corpus study on the use of

[24] On the basis of de Swaan's (2002; 2010) World Language System, Mair (2013) develops his World System of Standard and Non-Standard Englishes, which defines (Standard) American English as the *hyper-central variety* or *hub* of the World System of Englishes (264). The reason is that "other varieties are more likely to follow American usage than American English is to follow developments in other varieties" (Mair 2018: 51–52). Mair (2018) extends the model to a "non-colonial environment" (46), namely Germany.

uncountable nouns as countable nouns under the assumption that variety-specific, semantically motivated usage patterns (in the sense that certain uncountable nouns are relatively much affected in individual varieties) might emerge regarding that feature.

In order to allow for comparative analyses of the features of interest, corpora of similar or identical design were needed. Additionally, corpus size played a major role in the selection process because a usage-based account of the features considered requires databases big enough to investigate the impact of usage frequencies on omission and regularization rates. To the author's knowledge, ICE and GloWbE are the only existing corpora that combine parallel designs with sufficient amounts of data for the varieties investigated. Simplification as a deviation from standard use is expected to occur in spoken (unmonitored) language and in informal registers in particular, which determined the choice of corpus sections. Omission was primarily investigated in the spoken part of ICE, regularization exclusively in GloWbE.

Considering only the spoken sections of ICE left a corpus size of 600,000 tokens per variety, which proved to be sufficiently large to analyze omission but not regularization. The ICE project was initiated by the late Sidney Greenbaum (then Director of the Survey of English Usage, University College London) in 1988 and has seen the compilation of ICE corpora for various varieties of English since.[25] A common corpus design and annotation scheme allow for direct comparability of the different ICE corpora available. Each ICE corpus consists of 1 million words (600,000 words of spoken language and 400,000 words of written language). Table 4.1 summarizes the distribution of sections in the spoken part of ICE. Each text file contains approximately 2,000 words.

All the spoken sections of ICE were taken into account and random verb and noun samples were investigated therein instead of considering all verbs with past time reference and all nouns with plural references in face-to-face conversations, the most informal genre in ICE, only. This was necessary in order to account for usage frequency as a potential determinant of omission of inflectional marking in a small corpus and still be able to work with sufficient token frequencies. As sections 5.1 and 6.1 will show, past tense omission is to some extent phonologically conditioned, which required caution in the sampling process. Sections 5.2 and 6.2 discuss the respective sampling procedures in detail.

[25] The following ICE corpora are available: Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, Nigeria (written part), the Philippines, Singapore, Sri Lanka (written part), and USA (written part) (cf. Davies & Fuchs 2015: 2).

Table 4.1: Sections in the spoken part of ICE (The ICE Project 2016)

		section (number of text files)	file names
dialogues (180)	private (100)	face-to-face conversations (90)	S1A-001 to S1A-090
		phonecalls (10)	S1A-091 to S1A-100
	public (80)	classroom lessons (20)	S1B-001 to S1B-020
		broadcast discussions (20)	S1B-021 to S1B-040
		broadcast interviews (10)	S1B-041 to S1B-050
		parliamentary debates (10)	S1B-051 to S1B-060
		legal cross-examinations (10)	S1B-061 to S1B-070
		business transactions (10)	S1B-071 to S1B-080
monologues (120)	unscripted (70)	spontaneous commentaries (20)	S2A-001 to S2A-020
		unscripted speeches (30)	S2A-021 to S2A-050
		demonstrations (10)	S2A-051 to S2A-060
		legal presentations (10)	S2A-061 to S2A-070
	scripted (50)	broadcast news (20)	S2B-001 to S2B-020
		broadcast talks (20)	S2B-021 to S2B-040
		non-broadcast talks (10)	S2B-041 to S2B-050

Working with the Hong Kong component of ICE proved to be challenging insofar as a considerable number of speakers in the corpus are clearly not speakers of HKE. A look at the ICE-HK metadata showed that only 360 of the 647 speakers recorded for the corpus actually indicated to have been born in Hong Kong, although the number of speakers who reported to have received primary education in Hong Kong is higher (440). Luckily, the corpus marks extra-corpus speakers as "speaker Z." All utterances by speakers classified as extra-corpus speakers in ICE-HK were excluded from the corpus files prior to the analyses. While the extra-corpus speakers in ICE-HK could be handled easily, it is unfortunate that speakers other than HKE speakers served as speech partners in many of the conversations recorded. It remains unclear how much their way of speaking influenced that of the "true" HKE speakers.

The inflectionally marked and unmarked verbs and nouns were manually retrieved from the ICE corpora by means of the corpus analysis toolkit AntConc (Version 3.4.3m for Macintosh OS X; cf. Anthony 2014) on the basis of the plain .txt files. The part-of-speech tagged file versions (.pos) were opted against because the amount of data allowed for (less error-prone) manual analyses. The .pos files would not have made it possible to identify lack of inflectional marking more easily.

In addition to their analysis in ICE, the sampled verbs and nouns were investigated for lack of inflectional past tense and plural marking in GloWbE. GloWbE comprises 1.9 billion words of internet language from 20 countries[26] and makes a distinction between informal blogs (roughly 60 percent of the corpus) and other language material (called "general"; cf. Davies & Fuchs 2015: 2–3). This is supposed to resemble the spoken/written distribution (60 percent/40 percent) in ICE. In contrast with the ICE corpora, which contain data from the 1990s and 2000s, GloWbE was released in 2013 only and contains web documents collected between November and December 2012 (cf. Biber et al. 2015: 16). Table 4.2 shows the number of words, websites, and web pages per variety of interest and section. GloWbE IN (India) is about twice and GloWbE GB (Great Britain) and GloWbE US (United States) about nine times the size of GloWbE HK (Hong Kong) and GloWbE SG (Singapore).

Table 4.2: Sections in GloWbE by variety (Davies 2013)

corpus	section	words	websites	web pages
GloWbE SG	general	29,229,186	5,775	28,332
	blogs	13,711,412	4,255	17,127
	sum	42,974,705	8,339	45,459
GloWbE HK	general	27,906,879	6,720	27,896
	blogs	12,508,796	2,892	16,040
	sum	40,450,291	8,740	43,936
GloWbE IN	general	68,032,551	11,217	76,609
	blogs	28,310,511	9,289	37,156
	sum	96,430,888	18,618	113,765
GloWbE GB	general	255,672,390	39,254	232,428
	blogs	131,671,002	35,229	149,413
	sum	387,615,074	64,351	381,841
GloWbE US	general	253,536,242	43,249	168,771
	blogs	133,061,093	48,116	106,385
	sum	386,809,355	82,260	275,156

While the inflectionally marked verbs and nouns in GloWbE were retrieved on the basis of their part-of-speech tags, their unmarked counterparts had to be identified

[26]The 20 countries are the US, Canada, Great Britain, Ireland, Australia, New Zealand, India, Sri Lanka, Pakistan, Bangladesh, Singapore, Malaysia, Philippines, Hong Kong, South Africa, Nigeria, Ghana, Kenya, Tanzania, and Jamaica (cf. Davies & Fuchs 2015: 6).

manually. To be able to identify lack of inflectional marking in the large amount of data GloWbE offers, the sampling function the web-based interface of the corpus provides was used. The sampling function allows the user to choose between the display of 100, 200, 500, or 1,000 randomly sampled hits. In order to take account of the differences in corpus size across varieties, 200 hits per lemma were investigated for lack of inflectional marking in GloWbE HK and GloWbE SG and 400 in GloWbE IN, and normalized to corpus size. Initial searches had revealed no difference in omission rates in both the general and the blog sections, which is why both sections were considered.[27] Investigating omission in GloWbE allows one to examine to which degree the phenomenon has found its way into written compared to spoken language and to contribute to descriptions of the fuzzy category of internet language (see below). Because of the near absence of omission of inflectional marking in the Great Britain component of ICE, GloWbE GB was not considered (cf. sections 5.3 and 6.3).

For the regularization phenomena of interest, the spoken sections of ICE proved to be too small to make any meaningful observations, which is why those phenomena were investigated in GloWbE exclusively. Initial searches had shown that the general and blog sections are comparable in their regularization rates, so both sections were included in the analyses. Obviously, this approach does not test regularization in spoken language. Larger corpora comprising spoken language material are needed for that purpose instead. In contrast with lack of verbal past tense and nominal plural marking, the regularized target forms could be directly entered into the online search interface. Thus, it was not necessary to restrict the analyses to verb and noun samples. Verbs and nouns that are potentially prone to regularization were identified by means of lemma frequency lists retrieved from the full-text version of GloWbE.[28] Those lemma frequency lists served as approximations of the usage frequency of the relevant verb and noun lemmata in the varieties of interest. See sections 5.2, 6.2, 7.1.2, and 7.2.2 for details.

[27] In fact, 20 percent of the "general" searches in Google unavoidably provided the corpus compilers with blogs (cf. Davies 2013: 4). The blog section was compiled by means of targeted Google searches for blogs exclusively.

[28] This approach was made possible by a purchase of the full-text version of GloWbE by the DFG GRK 1624 "Frequency effects in language."

4.2 Using the web as a resource in corpus linguistics

The 2000s saw the emergence of the web as a new resource for corpus linguistic research. Gatto (2014: 36) refers to an innovative seminal paper by Kilgarriff from the early 2000s as one of the first sources that regarded the web as a promising linguistic resource at the turn of the millennium. The following lines are worth quoting here:

> The corpus resource for the 1990s was the BNC [British National Corpus]. Conceived in the 80s, completed in the mid 90s, it was hugely innovative and opened up myriad new research avenues for comparing different text types, sociolinguistics, empirical NLP, language teaching and lexicography.
> But now the web is with us, giving access to colossal quantities of text, of any number of varieties, at the click of a button, for free. While the BNC and other fixed corpora remain of huge value, it is the web that presents the most provocative questions about the nature of language (Kilgarriff 2001: 344, in Gatto 2014: 36).

As Gatto (2014) notes, "the way to treating the web as a linguistic corpus was by no means straightforward" (36), and the various contributions in Hundt et al.'s (2007a) volume *Corpus Linguistics and the Web* raise that issue in different contexts. One clear benefit of the web as a corpus is the sheer incredible amount of data the web offers. As Hundt et al. (2007b) point out, "carefully compiled corpora" (2) are simply too small to investigate many research questions of interest. On top of that, given the necessary skills and tools, web data can be retrieved relatively quickly, meaning that very recent data can be made available to the corpus linguistics community.

In order to get the URLs for the millions of web pages GloWbE comprises, the GloWbE compilers randomly searched for hundreds of high-frequency English 3-grams in Google, such as *and from the* and *and they are*, which were collected on the basis of the *Corpus of Contemporary American English* (COCA; cf. Davies 2008; Davies & Fuchs 2015: 4; Biber et al. 2015: 16–17).[29] The retrieved URLs were stored

[29] In order to minimize Google-internal search preferences, between 800 and 1,000 URLs were saved for each n-gram, which made sure that not only the topmost search results were retrieved. Compare Baroni et al. (2009) and Sharoff (2006: 17), in Biber et al. (2015: 17), for previous studies that used n-grams to manage web data.

in a database together with metadata like website, page title, and country. The association of website with country was straightforward for websites with a top-level country domain (".SG" for Singapore, for instance). With ".com" addresses, the compilers relied on information the Google "Advanced Search" limited by "Region" provided them with (e.g., IP address of the server, visitors to the website, top-level country domain of websites that link to the website). According to Davies & Fuchs (2015: 4–5), post-checking of hundreds of websites for their actual location proved the reliability of the country identification by Google. However, it has to be pointed out that a correctly identified country domain does not necessarily imply that the website contents stem from speakers of the respective variety of English (compare also Lüdeling et al. 2007: 15 for a word of caution on this issue). On the basis of the list of URLs gained, the web pages belonging to the retrieved websites were downloaded by means of HTTrack. "Boilerplate" material such as recurring headers and sidebars was removed with JusText, and the entire corpus was tagged by means of the CLAWS 7 tagger and imported into a database of the same architecture and interface other corpora from <corpus.byu.edu> have been imported to (cf. Davies & Fuchs 2015: 5).[30]

"Web as corpus" versus "Web for corpus building"

Before we continue with central benefits and pitfalls of using the web as a corpus, it makes sense to elaborate on different ways in which the web can actually function as a corpus. Hundt et al. (2007b: 2) make a distinction between "Web as corpus" and "Web for corpus building" (cf. de Schryver 2002; Fletcher 2004; 2007). While the former refers to uses of the web as a corpus itself that can be searched by means of commercial crawlers or search engines like WebCorp[31], the latter implies that web data are retrieved "for the compilation of large offline monitor corpora" (Hundt et al. 2007b: 2). GlobWbE is an example of the latter approach, which Hundt et al. (2007b) consider "methodologically somewhat safer" (3). They point out that when the web is used as a corpus itself "the machine is determining the results in a most 'unlinguistic' fashion over which we have little or no control" (3).

Gatto (2014) refers to "four basic ways of conceiving of the web as/for corpus" (37) established by Baroni & Bernardini (2006: 10–14): "The web as a corpus surrogate" functions as a resource for translation tasks or metasearches via WebCorp

[30] See https://www.httrack.com/, https://code.google.com/p/justext/, and http://ucrel.lancs.ac.uk/claws/.

[31] See http://www.webcorp.org.uk/live/.

and is an equivalent of the "Web as corpus" approach discussed in Hundt et al. 2007b: 2). Researchers using "the web as a corpus shop" (which is an equivalent of the "Web for corpus building" approach in Hundt et al. 2007b: 2) search and download web material by means of toolkits such as BootCat[32]. Researchers using "the web as a corpus *proper*" are interested in the web as a source that represents Web English. "The mega-Corpus mini-Web," finally, is described as "a new object (mini-Web/mega-Corpus) adapted to language research and combining web-derived (large, up-to-date, web-based interface) and corpus-like features (annotation, sophisticated queries, stability)" (Gatto 2014: 37). The following paragraphs adopt the somewhat broader distinction between "Web as corpus" and "Web for corpus building" used by Hundt et al. (2007b: 2).

Web registers

The web provides scholars with a range of new text types, as Hundt et al. (2007b) point out: "[A]part from email, there are chat-room discussions, text messaging, blogs, or interactive internet magazines—text types that are interesting objects of study in themselves" (1). From a more recent perspective, social media are worth mentioning as another highly dynamic and innovative text type, although n-gram searches such as those conducted by the GloWbE compilers do not retrieve social media contents because they are not freely accessible. Lüdeling et al. (2007) state that "[f]or the web as corpus, it is reasonable to assume that all categories of written language are represented to some extent" (14). Automatic Genre Identification, i.e., the use of computational methods to assign web texts to genres or registers by means of predefined descriptors, has had limited success so far because too little is known about "the full set of possible web registers" (Biber et al. 2015: 13) and their distribution.[33] Due to the enormous dynamism of the web, it is difficult or even impossible to go beyond a general distinction between prevalent and rare registers. Apart from that, inter-rater reliability between experts and lay users tends to be low (cf. Biber et al. 2015: 15; Rosso & Haas 2010). Those groups typically classify web documents manually on the basis of a list of register categories in Automatic Genre Identification studies.

According to Biber et al. (2015: 13–14), the following factors make it difficult to assign web texts to registers: Web texts often lack "external indication of register

[32] See http://bootcat.dipintra.it/.

[33] The authors use the term "register" instead of "genre" "to refer to situationally based textual distinctions on the web, following the research tradition developed in Biber 1995, Biber et al. 1999, and Biber & Conrad 2009" (Biber et al. 2015: 13).

category" (13), web documents are highly diverse (i.e., they lack common characteristics), and unpublished and published texts enjoy equal status and are difficult to distinguish. As mentioned above, GloWbE makes a broad distinction between "general" web material (20 percent of which unavoidably includes blogs) and blogs. The former were derived by means of "general" Google searches and the latter by means of "Google searches of just blogs" (Davies & Fuchs 2015: 4). As Mair (2015) puts it, "[w]hat the precise relationship is between informal digital literacy and actual spoken language is an extremely tricky issue, and so is the question whether blogs constitute a recognisable genre" (30–31).

In order to explore the composition of the searchable web, Biber et al. (2015: 17) use web pages gained by randomly extracting 53,424 URLs from GloWbE (US, UK, Canada, Australia, and New Zealand only) to "describe the lexico-grammatical characteristics of web documents" in the corpus. Instead of working with Automatic Genre Identification for the reasons mentioned above, the authors "developed a computational tool for register classification" (Biber et al. 2015: 18) implemented on the Amazon-based online crowd-sourcing utility Mechanical Turk[34]. 908 recruited raters coded the sampled web documents (four raters per document) for basic situational characteristics rather than the register category itself, namely "the mode (spoken or written), relations among participants (multiple interacting participants *versus* authors who do not interact with addressees), and communicative purposes (e.g., to narrate, to inform, to express opinion)" (Biber et al. 2015: 19; emphasis in the original). On the basis of these choices, raters chose general registers (e.g., narrative) and specific sub-registers (e.g., personal blog; for an overview of this hierarchical framework see ibid.: 21, table 1). For about 69 percent of the web pages sampled, at least three raters agreed on the same general register, and 29 percent of the web pages constitute "hybrid" registers (e.g., Narration+Opinion). A few general registers and hybrid combinations dominated the registers identified, such as news/sports reports/blogs (about 21 percent of the web documents), informational descriptions/explanations (about 14 percent), and opinionated texts (about eleven percent; ibid.: 40). Currently, the authors are working out category-specific lexico-grammatical characteristics with the aim to "document systematic linguistic patterns of register variation on the web" (ibid.: 41). Their findings speak in favor of the applicability of a hierarchical framework for rating web registers, but the fact that 29 percent of the web pages were assigned "hybrid" registers poses a challenge

[34]See https://www.mturk.com/mturk/welcome.

for established register distinctions. Additionally, the lack of clear register distinctions makes it difficult to arrive at a balanced corpus, i.e., one that "cover[s] a wide range of text categories considered to be representative of the language or variety under scrutiny" (Gatto 2014: 12).

Representativeness

Besides corpus size, corpus compilation, and register variation in web-based corpora, a few words should be said about two further prerequisites a corpus is supposed to meet, namely representativeness and reproducibility. The issue of representativeness ties directly in with the range of new text types the web offers and with the problem of register definition when it is not clear which registers to expect (compare Biber et al. 2015 above). Leech (2007) points out that "[w]ithout representativeness, whatever is found to be true of a corpus, is simply true of that corpus—and cannot be extended to anything else" (135). While compiling a representative corpus is an ambitious goal in general, the important question is whether this task is particularly difficult for compilers of web corpora. The somewhat fuzzy register categories and the transient nature of (many) web contents make it certainly difficult to determine what exactly a web-based corpus is supposed to represent.

Reproducibility

According to Gatto (2014), "[o]ne of the most obvious practical consequences for linguistic research of the web's dynamic nature is the impossibility of reproducing any experiment" (68). Lüdeling et al. (2007: 10) point out that, ideally, corpus results should be replicable when conducting the same analyses on a parallel corpus that fulfills the same criteria catalogue. The problem with using the web as a corpus ("Web as corpus") is that the web is constantly changing with web pages being added, updated, or deleted. The issue is particularly serious when a commercial search engine is used that "employs algorithms which are totally mysterious to the average user" (Leech 2007: 144). In fact, "paid positioning," which "is intended to steer searchers away from more relevant 'natural' search results toward advertisers' sites" (Fletcher 2007: 30) strongly influences the search results we are presented with. Thus, we do not know about the database or search strategies commercial search engines are working with (cf. Lüdeling et al. 2007: 11). As regards reproducibility, working with corpora (like GloWbE) that consist of websites that were retrieved from the web in a controlled manner to create a fixed database ("Web for corpus building") is definitely the safer option.

Precision and recall

Two further concepts addressed here are precision (relevance of the search results) and recall (reliability of the search results). According to Gatto (2014), with web data "recall is impaired by its 'unstable' nature as a dynamic non-linguistically oriented collection of text, whereas precision is impaired by the intrinsic limitations, from the linguist's perspective, of search tools such as ordinary search engines" (69). Irrespective of corpus type, low recall (many correct items, i.e., *false negatives*, are missed) is highly problematic for both quantitative and qualitative analyses, whereas low precision (many wrong items, i.e., *false positives*, are returned) can be met by going through the results manually as long as this is technically and practically feasible (cf. Lüdeling et al. 2007: 12). When a corpus is annotated, the quality of the retrieved items additionally depends on the quality of the linguistic annotations. With the "Web for corpus building" approach, precision and recall can be handled as with any other corpus type, although it should be mentioned that unedited web material such as comments or forum entries tend to contain misspellings that part-of-speech taggers cannot deal with. Additionally, features that require low precision searches confront the researcher with huge amounts of hits that need to be gone through manually (compare section 4.1 for the samples drawn from GloWbE to investigate the omission phenomena of interest). Duplicates are another issue, which the GloWbE compilers dealt with by only allowing each web page to enter the list of web pages retrieved by Google searches once, by removing boilerplate material like headers and sidebars (which reoccur on web pages belonging to the same website), and by searching for duplicate n-grams (primarily 11-grams) to check for repetitions of long strings of words.[35] The Keyword in Context (KWIC) display additionally indicates all duplicates that have been identified subsequently and logged in the database by providing the number of duplicates in brackets after the respective web page(s). These duplicates are being eliminated from the database in regular intervals.

When the web is used as a corpus that is searched by means of search engines, precision, recall, and duplicates are far more serious issues. Firstly, search engines use normalization such as case insensitivity and automatic variant identification (searching for "white space" returns *white space*, *white-space*, and *whitespace*) that cannot be easily deactivated (cf. Lüdeling et al. 2007: 14). Secondly, duplicates cannot be controlled for apart from checking them manually, which is complicated

[35] See http://corpus.byu.edu/glowbe/ → Texts and registers → Notes on duplicate texts.

because all web pages need to be downloaded and might not even be traceable in case the search engine poses a limit on the number of displayed hits.

In sum, the web is a promising new resource for corpus linguistic research because its size and speedy compilation allow for answering new and pressing research questions. Much still needs to be learned about register distribution before issues like representativeness and balancing can be considered to sufficient degrees though. The general impression gained is that challenges regarding reproducibility as well as precision and recall (including the handling of duplicates) arise with the "Web as corpus" approach and the reliance on (commercial) search engines in particular. In line with Hundt et al. (2007b), the "Web for corpus building" approach can be considered as methodologically safer, although the sheer size of GloWbE proved challenging for investigating lack of inflectional marking where only small samples could be accounted for. For the regularization features of interest with much fewer *false positives*, however, GloWbE proved to be a very convenient tool.

4.3 Conducting a web-based experiment

In addition to the corpus analyses, a web-based experiment was conducted in order to test the perception of lack of verbal past tense and nominal plural marking among speakers of English from Hong Kong, India, and Singapore (chapter 8). Speakers of BrE and AmE served as the control group. Both BrE and AmE were designated as control varieties because they do not differ in their past tense and plural marking systems. The experiment consisted of two tasks, namely a self-paced reading task and an acceptability judgment task, which tested two fundamentally different aspects. The self-paced reading task was performance-based in that participants read stimuli (i.e., sentences) word by word at their own pace by pressing a keyboard button or touching the screen to get from one word to the next. The reading times were measured as reaction times in between pressing the button or touching the screen. In the acceptability judgment task, participants judged sentences on a scale from "not acceptable at all" to "fully acceptable." Instead of testing the participants' immediate reaction to unmarked forms, this task aimed at their metapragmatic assessment of the features of interest, which is likely influenced by familiarity with the features and language ideologies.

The main reasoning behind combining investigations of production (corpus analyses) and perception (experiment) data was to elaborate on the "usage despite aware-

ness" phenomenon that has been described for SgE in particular. The local vernacular Singlish functions as an identity carrier among Singaporeans and is used in informal domains despite governmental efforts to limit its use (compare section 3.2). The mixed-methods approach adopted here allows investigation of both usage patterns and their assessment under the assumption that the more omission occurs in a variety, the faster unmarked verbs and nouns are read, although they are not (necessarily) evaluated more positively than by speakers who are less familiar with omission (i.e., the control group). Details on the design of the self-paced reading task and of the acceptability judgment task will be provided in section 8.2.2. In the following paragraphs, the focus lies on the advantages and pitfalls of conducting an experiment in a web-based manner and on the particular challenges this poses to measuring reaction times.

Advantages and pitfalls

One clear advantage of web-based experiments is the seemingly immediate availability of large numbers of participants. As Crump et al. (2013) put it: "In theory, online experimentation would allow researchers to [sic] access to a large and diverse pool of potential subjects worldwide, using automated replicable techniques free of unintended experimenter effects" (1). In fact, it was only by means of a web-based experiment that all three target groups of interest in this book as well as the control group could be reached in a manageable amount of time. Crump et al. (2013: 1) continue by pointing out that the challenges come with actually finding people on the internet who are willing to participate and with compensating them. Online crowdsourcing services such as Amazon Mechanical Turk (AMT) and Prolific[36], where online users can register to anonymously participate in web-based surveys for small monetary rewards, are an appealing alternative to the traditional recruiting of "university undergraduates who participate in studies in exchange for experience, course credit, or money" (ibid.: 1). Unfortunately, AMT and Prolific do not attract sufficient numbers of potential participants from Hong Kong and Singapore, which is why those crowdsourcing services proved insufficient for the purposes of this book. Among the roughly 50,000 registered users of Prolific that provided their country of birth, about 60 percent were born in the US and the UK (cf. Prolific 2016). Ipeirotis (2010) reports for AMT that about 50 percent of the registers users (labeled "workers" in AMT) are from the US and about 40 percent from India.

[36] See https://www.prolific.ac/.

Participants were recruited by means of the so-called "friend-of-a-friend" approach, i.e., friends were contacted who then shared the call for participation with their friends. The friend-of-a-friend approach is based on the assumption that a group can be reached best by contacting group insiders or friends of group insiders (Milroy 1980; Milroy & Gordon 2003: 73–76). The call for participation was shared via the social media platform Facebook and via email. Participants had the possibility of entering a raffle to win one of 30 payments of 15 euros converted to their currency and transferred via PayPal, which made web-based compensation feasible.

Traditional lab settings are tightly controlled by the experimenter (e.g., same lab with identical device(s) used, millisecond timing of stimulus presentation and response recording, attention and commitment on part of the participants, possibility of clarifying questions right away, which allows for more complex instructions; cf. Crump et al. 2013: 1). In contrast, web-based experiments save resources because they do not require lab preparation, and they are time-effective because different participants can take part simultaneously and do not need to be monitored (cf. Enochson & Culbertson 2015: 1). However, the experimenter can neither control the setting nor can the participants' commitment be expected to be as strong as if the experimenter were present. Connected to that, Crump et al. (2013) stress "two key challenges for online data collection" (1), namely technical challenges (such as guaranteeing that all necessary features are supported by various browser systems) and the particular sensitivity of tasks where timing is crucial to the testing environment. Enochson & Culbertson (2015: 13) point out that response time measurements pose a particular challenge because they can differ across devices, for instance because of different keyboards or differences in the refresh rates of monitors.

The validity of measuring reaction times web-based

To get a better idea of the degree to which reaction times differ between data collected in the lab and web-based data, several authors have replicated experiments web-based that had previously been conducted under laboratory settings. Using the web-based experimental software WebExp[37], Keller et al. (2009) replicate self-paced reading results on parsing ambiguity. Enochson & Culbertson (2015) reproduce the faster processing of pronouns compared with determiner phrases, the processing costs of filler gaps in *wh*-fronted constructions, and agreement attraction. In both studies, millisecond response time data were collected. Research has shown that using JavaScript to record response times client-side (i.e., locally on the computer

[37]See http://groups.inf.ed.ac.uk/webexp/.

of the participant) and collating them on AMT upon task completion "is precise enough to capture a number of classic effects in cognitive psychology [...] even when the tasks require sustained attention and complex instructions" (Enochson & Culbertson 2015: 2). Enochson & Culbertson (2015: 2–3) use ScriptingRT[38], a Flash-based open-source software that visually displays words or images with millisecond precision and records response times with the same precision. With diminishing support for Flash by Mozilla, Safari, and Chrome lately, the implementation of ScriptingRT in the experiment software may require participants to install Flash in order to be able to do the online experiment. This is a major drawback that speaks against the use of Flash-based software.

A Python-based software was used for the experiment presented in chapter 8 that was explicitly programmed for the purposes of the experiment and that allows implementation of the required designs for both the self-paced reading task and the acceptability task (see section 8.2.2 for details).[39] All data points were saved in an SQL-based database that makes it possible to dynamically retrieve purpose-specific dataframes without ending up with one huge dataframe that is difficult to handle. Existing open-source experimental software like WebExp, Ibex, or DMDX (MS Word-based) lacks this latter function; a function that proved to be very convenient for the analyses. The database also allows for the direct implementation of frequency lists, syllable lists, and the like. The web-based experiment works across platforms and browsers, and both platform and browser were logged in the database. The consent form preceding the experiment informed participants about all information collected and saved in the database.

Response times were recorded client-side using JavaScript in order display the stimuli just in time and to precisely record the reaction times. Additionally, dynamic updates were implemented, meaning that data were continuously sent to the server. In case the internet connection broke or participants (accidentally) closed the browser tab the experiment ran in, participants were redirected to the last page they had seen. This was realized by means of session cookies implemented primarily to avoid that participants took part multiple times. Every page reload (including the redirection to the last page seen) was logged in the database. Additionally, all clicks and button presses were timestamped in order to identify pauses. Reloaded

[38] See https://reactiontimes.wordpress.com/2016/10/15/end-of-flash-is-end-of-scriptingrt/.

[39] Many thanks to Martin Isack for programming the software. The software is open source and is made available upon request. Please contact the author via research@terassa.de.

stimuli and the first two stimuli that directly followed a pause were excluded from the analyses.

4.4 Quantitative and qualitative analyses

Quantitative and qualitative methods were combined to analyze the results of the corpus studies and the experiment presented in the following chapters. Approaching the features of interest from a quantitative perspective helped to get an overview of the underlying omission and regularization patterns. The usage patterns identified in the quantitative analyses were analyzed in more detail in a second step. Qualitative analyses also preceded the quantitative analyses in the corpus studies on omission of verbal past tense and nominal plural marking, where potential hits had to be retrieved and coded manually.

4.4.1 Quantitative analyses

The corpus results were analyzed by means of descriptive statistics for the simple reason that the number of verbs and nouns that lack inflectional past tense and plural marking and that are regularized were too low to opt for significance testing. In the corpus studies on past tense and plural omission, a number of factors that potentially impact on omission rates were controlled for, such as the morphological process applied to form the past tense of verbs, syllable number (verbs and nouns), the presence of time adverbials (verbs) and plural determiners like quantifiers or numerals (nouns), genre-specificity (verbs and nouns), and usage frequency (verbs and nouns).

Working with frequencies of use

Both in the corpus studies and in the experiment, frequencies of use were approximated by means of lemma token frequencies. The only exception is a comparative analysis of omission of verbal past tense marking by the morphological process involved in marking past tense inflectionally. Here lemma type frequencies were accounted for (see section 5.4, and in particular figure 5.6 therein). For the corpus analyses, lists of the lemma token frequencies of all verbs and nouns in the GloWbE corpora of interest were extracted from the GloWbE full-text offline database (see above). Relative rather than absolute frequencies were chosen to account for differences in corpus size (see table 4.2). To arrive at variety-specific relative lemma token frequencies, the variety-specific absolute lemma token frequencies derived from the

GloWbE full-text offline database were divided by the number of words in the respective GloWbE corpus. For further analyses, the relative lemma token frequencies were logarithmically transformed. Logarithmic transformation is commonly used to reduce the amount of skewing in the distribution of variables (e.g., Baayen 2012: 31). For many, if not most, statistical techniques, it is necessary to control for extreme outliers that would otherwise change the overall picture drastically. Additionally, logarithmic transformation is useful for data visualization because it makes data points spread more evenly across the graph rather than having them cluster tightly in densely populated frequency regions. For the mechanisms behind logarithmic transformation, see Field et al. (2012: chapter 5).

Regression modeling

The experimental data were analyzed by means of regression modeling. Linear mixed-effects models were used. Multifactorial methods, in general, are a valuable tool for taking account of the impact of multiple independent variables on the dependent variable (cf. Gries 2013: 239). Mixed-effects models, in particular, go one step further by controlling for both fixed and random effects. Let us briefly look at the `lexdec3`-example provided by Baayen (2012: 244–246) to grasp what is meant by that. Baayen (2012: 244) fits a mixed-effects model to test the effect of familiarization with a task on reaction times (`RT`) and measures the degree of familiarization by the position of the trial in the task (`Trial`). This is the model formula:

```
lexdec3.lmer = lmer(RT ~ Trial + (1|Subject) + (1|Word), lexdec3)
```

Besides the fixed effect `Trial`, the formula contains two random effects, namely the random intercepts `(1|Subject)` and `(1|Word)`. Fixed effects are effects all kinds of regression models account for, irrespective of whether the models are mono- or multifactorial (see above). Random effects, which are limited to mixed-effects models, allow making adjustments to better account for variances in variables. In the `lexdec3`-example above, the random intercepts `(1|Subject)` and `(1|Word)` make sure that the intercept[40] is adjusted for subjects who are particularly quick or slow responders and for words that are relatively frequent or familiar, infrequent or little familiar, or otherwise specific. With corpus data, it can make sense to adjust for lemma-specific effects. Figure 4.1 plots `RT` as a function of `Trial` by subject for the data set `lexdec3`.

[40]The intercept is the Y-coordinate where the regression line crosses the Y-axis (Baayen 2012: 85).

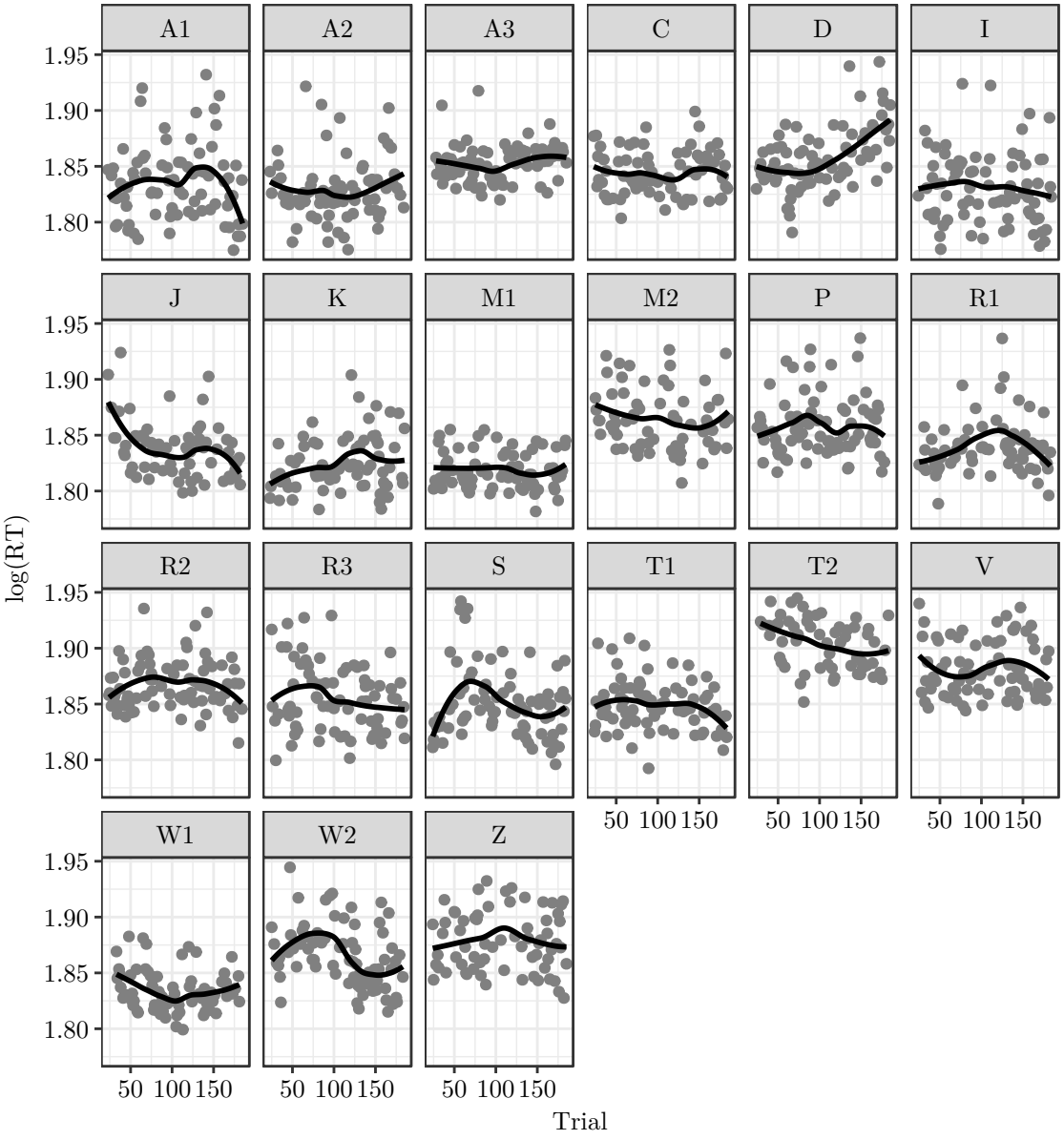

Figure 4.1: RT as a function of trial by subject (data set "lexdec3," locally weighted smoothing (called "loess") applied; adopted from Baayen 2012: 245 and adjusted)

Locally weighted smoothing called "loess" (acronym for "local regression"; e.g., Jacoby 2000) was used to depict the regression lines (in black in figure 4.1). Participant M1, for instance, is a quick responder with no sign of fatigue, which would show in case reaction times increased with later trials, as it is the case with participant D (cf. Baayen 2012: 244). Adding `(1|Subject)` as a random intercept adjusts the reaction times of participant M1 "for the average speed by means of small changes to the intercept [of participant M1]" (ibid.: 245).

As pointed out above, the model formula contains one fixed effect, `Trial`, and two random intercepts, `(1|Subject)` and `(1|Word)`. One could add further predictor variables as fixed effects. Depending on whether their effect is additive or whether they interact with the already present fixed effect, the model formula needs to be adjusted accordingly. Two independent variables are "*additive* [...] when the combination of the two variables has the effect [...] expect[ed] on the basis of each variable's individual effect" (Gries 2013: 20; emphasis in the original). In contrast, "[t]wo or more variables interact if their joint effect on the dependent variable is not predictable from their individual effects on the same dependent variable" (Gries 2013: 21). Let us look at an example provided by Gries (2013: 20–22) to make this clear. The example studies on the basis of a fictitious data set whether the length of a clause constituent (`Length`) depends on the grammatical role the constituent plays (`GrmRelation`; levels: subject, object) and on the clause type (`Clause type`; levels: main, subordinate). The two interaction plots in figure 4.2 depict an additive effect (plot to the left) and an interaction (plot to the right). Length is measured in syllables.

The plot to the left shows that subjects as well as constituents of main clauses are relatively short. The additive effect is visible insofar as subjects that occur in main clauses are shortest, whereas objects that occur in subordinate clauses are longest (cf. Gries 2013: 20). The plot to the right also reveals that subjects and constituents of main clauses are relatively short. However, the additive effect is missing because objects in subordinate clauses turn out to be shorter than subjects in that same clause type (ibid.: 21). The interaction of `GrmRelation` and `Clause type` is clearly visible because the slopes of the lines have different signs (the lines even intersect in the example). The slope of a regression line specifies "how far [one] ha[s] to move along the horizontal axis for a unit change in the vertical direction" (Baayen 2012: 85). The interaction effect could also be less obvious in case main clause objects are

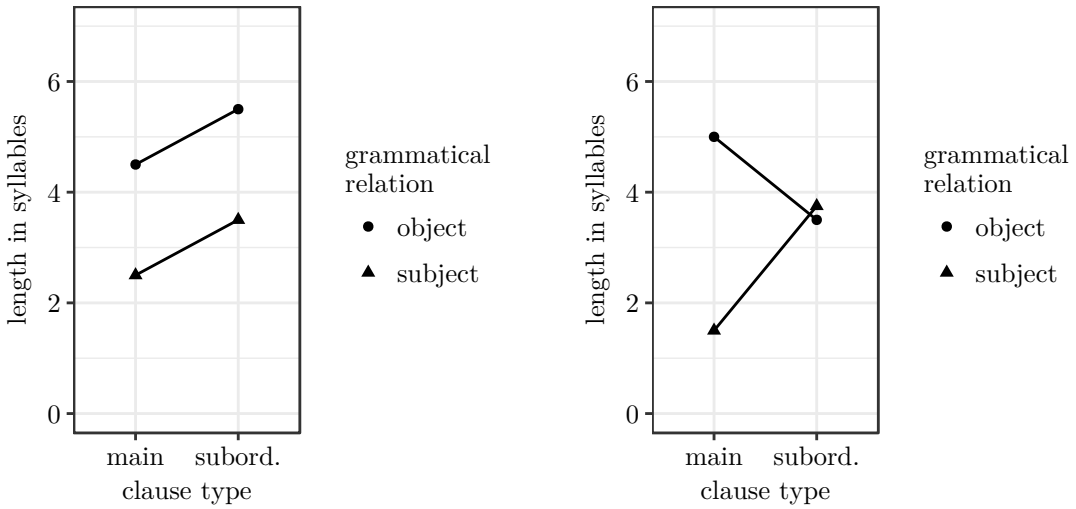

Figure 4.2: Interaction plots with additive effect (left) and interaction effect (right) (adopted from Gries 2013: 21–22 and adjusted)

only slightly longer than main clause subjects but subordinate clause objects are considerably longer than subordinate clause subjects (cf. Gries 2013: 22).

For further details on regression modeling in general and mixed-effects regression modeling in particular, the reader is referred to introductory works on statistics in linguistics such as Gries (2013: chapter 6) and Baayen (2012: chapters 6 and 7). Another highly recommended comprehensive introduction to statistics from the field of psychology is the one by Field et al. (2012) mentioned above. All those works focus on analyzing data by means of R (The R Foundation 2017), which is a programing language and environment for statistical analyses that has enjoyed increased popularity in linguistics. All analyses conducted for this book were carried out by means of R. The latest version used was version 3.3.3, released on 6 March 2017. For the analyses, R was run in RStudio (RStudio 2016), an open-source integrated development environment for R. Here, the latest version used was version 1.0.136, released on 21 December 2016. The package "lme4" (Bates et al. 2017) was used to fit the regression models.

4.4.2 Qualitative analyses

To prepare the corpus data for the quantitative analyses, the inflectionally unmarked verbs and nouns as well as their marked counterparts were manually retrieved from ICE. Besides lemma-specific characteristics such as syllable number and lemma token frequency, all individual hits were coded for variables like the presence or absence of

a time adverbial with past time reference (verbs) or the presence of a determiner like a quantifier or numeral with plural reference (nouns), perfectivity (verbs only), and the sound following the inflectional suffix (verbs only), among others. Instances of self-correction were not counted as hits. The manual coding helped get a feeling for the corpus data and for the context the features of interest occur in. Corpus analyses can be conducted in different ways. For instance, a certain amount of corpus files can be read and investigated manually or, to mention the other extreme, large amounts of hits can be retrieved automatically in case the features investigated allow for automatic retrieval. The latter approach was adopted for the regularization features of interest. For the reasons mentioned in section 4.1, manual coding was necessary to identify verbs and nouns that lack inflectional past tense and plural marking, respectively. The usage patterns the quantitative analyses revealed were then analyzed in more detail, among other things by accounting for substratum transfer as a potential explanatory factor.

5 Omission of inflectional past tense marking: A corpus-based account

This chapter presents a corpus study on omission of inflectional past tense marking in the three Asian varieties of interest. Both regular and irregular verbs are considered. Lack of inflectional past tense marking in the presence and in the absence of a time adverbial is accounted for. Compare examples 5.1 (no time adverbial) and 5.2 (time adverbial), respectively, where *show* and *stay* lack their past tense *-ed* suffix.

(5.1) Uhm you know he *show* uh compassion uh and it was <unc> one-word </unc> at the you know at at the scene (ICE-HK:S1B-036#96:1:B)

(5.2) He *stay* in uh Temasek Hall last time (ICE-SIN:S1A-077#129:1:B)

A particular focus lies on the potential and limitations a usage-based account of omission of inflectional past tense marking offers. Thus, the impact of frequencies of use (approximated by the lemma token frequencies of a sample of verbs) on degrees of omission of inflectional past tense marking is a main point of interest. This aspect has not received much attention in the World Englishes literature. ICE and GloWbE served as databases that allow for contrastive analyses across the varieties of interest.

Besides frequencies of use, two further factors that potentially impact omission rates are taken into account, namely substratum influence and the degree of institutionalization of the contact varieties. The phonetic environment is particularly focused on because consonant cluster reduction can affect the inflectional suffix. As mentioned in section 4.1, a look at omission rates in GloWbE give insights whether omission has found its way into (arguably informal) language use on the web, thereby contributing to attempts at characterizing internet language.

Sections 5.1 and 5.2 provide an introduction into previous research on omission of inflectional past tense marking and present the sample of verbs chosen and the hypothesis underlying the corpus study, respectively. Section 5.3 gives a general overview of the observed omission rates, whereas section 5.4 offers a usage-based account of omission of inflectional past tense marking in ICE. Section 5.5 investigates

the potential impact of substratum transfer and institutionalization on omission rates and elaborates on whether those factors constrain potential frequency effects, before section 5.6 concludes.

5.1 State of the art

Omission of inflectional past tense marking is a feature that is well attested for varieties of English around the world. According to eWAVE, feature 132 "zero past tense forms of regular verbs" is pervasive or obligatory in varieties of English across Asia, Australia, Africa, and America, and it is one of the L2-simple features Kortmann & Wolk (2012) identify. However, it is neither among the top L2 nor among the top Asian features (compare section 2.1, table 2.2). It received A-ratings (pervasive or obligatory feature) for HKE and CollSgE and a C-rating (extremely rare feature) for IndE.

Singapore English

The literature frequently mentions omission of inflectional past tense marking as a typical feature of SgE and has investigated the phenomenon from different angles. The absence of overt marking for past tense is often explained by influence from substrate languages like Mandarin, Hokkien, or Malay that are acquired as first languages in Singapore (e.g., Platt & Weber 1980; Yeo & Deterding 2003). M. L. Ho & Platt (1993) point out that speakers of Chinese languages are faced with "considerable problems" (74) when acquiring the English tense-aspect system. Chinese lacks verbal tense marking and is aspect-prominent in contrast with tense-prominent English. Systematic evidence of substratum transfer as a determinant of past tense omission in SgE as a result of that tense-aspect difference is provided by Sharma (2009: 177–179). On the basis of data by M. L. Ho & Platt (1993), she reports that the majority of the verbs in perfective contexts (56.2 percent, N=8,725) are overtly marked for past tense, compared with 14.7 percent in imperfective contexts with habitual meaning and 36.9 percent in imperfective contexts with stative meaning.[41] Sharma (2009) interprets this as a sign of "direct replication [...] of perfectivity marking in the substrate system[ ] [of SgE]" (177–179). Mandarin explicitly "indicates perfective aspect and relative past time reference" (Comrie 1976: 58, in M. L. Ho & Platt 1993: 74) by means of the verbal particle *le*. Sharma (2009) concludes

[41] Ho and Platt's (1993) data stem from conversational interviews conducted among 100 ethnically Chinese Singaporeans.

that "perfective meaning [is clearly ascribed] to English past morphology" (176), which is why verbs in perfective contexts tend to be overtly marked and verbs in imperfective contexts not.

With reference to Bao (1995; 2005), Sharma (2009) additionally points out that speakers of SgE often use *already* as a perfective marker, following "formal analogy with analytic Chinese and Malay forms *le* and *sudah*" (179). According to Bao (1995: 183), *already* is used both in perfective contexts (completion of an action, e.g., *I work about four months already*) and in inchoative contexts (onset of an action, e.g., *My son go to school already*); the former being motivated by lack of verbal tense marking.[42] He assumes that lack of verbal tense marking also accounts for uses such as *Everybody down there see me before* (cf. Tay 1979: 104), where *before* indicates the past time reference. In a similar vein, Bao (1998) and Alsagoff (2001) point out that in casual speech time adverbials like *yesterday* provide temporal reference while the verb remains unmarked for tense. Uses of "narrative present" are worth mentioning in that context, where present tense is used once the past tense time frame has been set up; usually by marking the first verb(s) for past tense and switching to present tense then (e.g., M. L. Ho 2003; Deterding 2007: 46–47; Gut 2009b: 273–274).

Besides substrate influence, two other determinants of past tense omission are commonly discussed in the literature, namely the syllable structure, i.e., the sound(s) preceding the past tense suffix, and verb meaning (cf. Silver et al. 2009: 135). Let us turn to the phonetic environment the past tense suffix is embedded in first. Consonant cluster reduction is a typical feature of SgE and affects past tense omission in regular verbs, when the verb ends in a consonant other than a plosive that together with the past tense suffix *-ed* adds up to a consonant cluster. Irregular verbs whose past tense is formed by means of vowel change and the addition of an alveolar plosive (e.g., *think*, *mean*, or *tell*), in contrast, retain their past tense marking because of the changed vowel even if the final plosive is lacking. Platt & Weber (1980: 59–61) observe that verbs ending in a consonant are more likely to lack past tense marking than verbs ending in a vowel. Randall (1997; 2003) and Yip (2004, both in Silver et al. 2009: 136) find similar patterns in the acquisition of past tense marking by Singaporean children and students. In Randall's (2003) data, "88 % of spelling errors made by students involved the omission of the final consonant in words whose final cluster consisted of a suffix" (Silver et al. 2009: 136), which he attributes to

[42] Bao (2005: 242) discusses a third use of Chinese *le*, namely with inceptive meaning, as in *wǒmen chī̄ le liúlián* ("We started/are about to eat durian").

the lack of word-final consonant clusters in both Chinese and Malay (cf. Randall 2003: 3). Deterding (2007: 41, 46) refrains from providing a quantitative account of past tense omission because of the possibility that past tense omission is a result of final plosive deletion in consonant clusters. Regarding lack of past tense marking in irregular verbs, he reports the following instead: When analyzing the recorded speech of a female university undergraduate of ethnically Chinese origin, Deterding (cf. 2007: 6–7) notes that the subject starts with a verb in the past tense and then switches to the present. In some cases, the subject seems to refer to something that is still true of the present, whereas in other cases use of narrative present prevails. More precisely, once the time frame is set, the subject switches to the present tense (compare also M. L. Ho 2003).

Based on the original recordings collected for the *National Institute of English Corpus of Spoken Singapore English* (NIECSSE; cf. Deterding & Low 2001), Gut (2009b) observes that the rather formal style prevalent in the NIECSSE is marked by a "largely functioning morphological tense marking system" (272). Only 20 percent of the verbs (both regular and irregular) in a past tense context lack verbal past tense marking, and verbs that form their past tense by means of /t,d/ affixation are comparatively highly affected by omission (ibid.: 267).[43] Gut (2009b) concludes that omission in SgE is predominantly phonologically conditioned because mainly forms "that do not run counter to the more dominant phonological processes of [cluster-]final /-t,d/ deletion" (273) mark the past tense inflectionally. Unmarked irregular verbs likely constitute present tense forms instead. Gut (2009b: 273) reports that 76.6 percent of the unmarked forms that cannot be a result of final plosive deletion are used when habitual actions or actions in the past with relevance for the present are described, when a time adverbial or "expression of the past" (ibid.) is present, or when the past time reference has been set up by a preceding marked verb or by a time adverbial. She also compares retention rates of cluster-final /-t,d/ and single final /t,d/ in verbs where the final plosive represents the inflectional past tense marker with respective retention rates in other lexical words, and observes that both single final as well as cluster-final alveolar plosives are retained more in past tense suffixes than in other lexical words (ibid.: 274). She assumes that particularly "highly

[43] /t,d/ affixation comprises [t], [d], and [ɪd] affixation. In Gut's (2009b) study, only 40.7 percent of the verbs forming their past tense by means of [t] or [d] affixation and 41.7 percent of those with [ɪd] affixation are marked for past tense compared with 90 percent of the verbs that form their past tense by means of suppletion, 77.4 percent of the verbs with vowel change, and 50 percent of those with vowel change plus [t,d].

educated speakers of SgE" (ibid.: 275) tend to actively suppress /t,d/ deletion when the past tense suffix is affected. Fong (2004) claims that when a time adverbial with past time reference occurs in sentences with third person singular subjects and the verb is unmarked, the verb is non-finite because agreement marking is lacking. In contrast, when agreement marking is not an issue, we cannot tell whether the unmarked verb formally represents a finite or a non-finite form.

The third factor often mentioned in the literature on past tense omission in SgE besides substrate influence and the influence of the phonetic environment, is verb meaning. M. L. Ho & Platt (1993: 40) note that stative verbs (i.e., verbs that refer to a state) and non-punctual verbs (i.e., verbs that refer to an habitual action) are more likely to lack past tense marking than punctual verbs (i.e., verbs that refer to a completed action). Compared with 36.9 percent of the stative verbs and 14.7 percent of the non-punctual verbs, 56.2 percent of the punctual verbs are marked in their study. Saravanan (1989, in Silver et al. 2009: 136) observes the same pattern among "Tamil speakers of Singapore English" (Silver et al. 2009: 136).

Silver et al. (2009: 134–135) present the so-called Aspect Hypothesis and test it for SgE student writing. The Aspect Hypothesis "postulates that the (lexical) aspectual meaning of verbs—the ways in which verbs describe the completion and duration of events—affects the degree to which they are accurately marked for tense and grammatical aspect" (ibid.: 135). It predicts that telic verbs (i.e., verbs with a natural endpoint such as *paint* or *bake*) have a more developed past tense marking system (i.e., are mastered earlier) than non-telic verbs (i.e., verbs without a natural endpoint that describe states or activities such as *love* or *enjoy*). The study is based on previous research by Yip (2004) and Yap (2006). While Yap (2006) claims that patterns of past tense use in secondary school student writings can be explained with the Aspect Hypothesis, Yip (2004) observes that syllable morpho-phonological constraints lead to higher error rates in student essays than lexical aspect. Silver et al. (2009) find higher error rates in the past tense marking of non-telic verbs than in that of telic verbs, both in primary school student writings (ibid.: 141) and in secondary school student writings (ibid.: 142).[44]

[44]The authors are aware of Ellis' (1994: 74) note that error distributions do not allow for more than cautious conclusions about language acquisition (cf. Silver et al. 2009: 137). Errors are defined as deviances in usage from standard BrE in their study (cf. ibid.: 144, footnote 1).

Hong Kong English

Research interest on past tense marking in HKE is growing. In line with McArthur (2002: 360), who claims that the present tense tends to be used in HKE irrespective of time reference, Setter et al. (2010) notice that "speakers do not show a great variety of verb tenses in their speech" (49) in that they mainly stick to present and past tenses.[45] The authors provide examples of tense switching in narrations where some verbs are marked for past tense and others are not, although speakers tend to switch back and forth rather than to stick to the present tense once the past time reference has been established (compare uses of "narrative present" in SgE; see above).

Wong (2017) stresses that the "marked difference between Cantonese and English with regard to tense [in the sense that Sinitic languages do not distinguish between past and present whereas English does] explains why tense contrasts are suspended in HKE, where a verb is clearly used with past time reference but appears in the base form" (15). Instead, Cantonese marks time "by a combination of adverbials, aspect markers and contextual factors" (ibid.: 16, with reference to Matthews & Yip 1994: 198). She concludes that substrate transfer is the driving force behind specifying the time frame by means of adverbials in HKE.

Gisborne (2009) raises the issue of tense marking in a study on the expression of finiteness in HKE. Following a discussion of Mandarin by Hu et al. (2001), he argues that Cantonese (like Mandarin) lacks finiteness, which "in the western European tradition [..] is associated with tense marking and verbal inflection, verb-subject agreement and the requirement of clauses to have subjects" (Gisborne 2009: 154; compare also Nikolaeva 2007).[46] In English, the matrix verb determines whether a finite or a non-finite complement clause follows (compare *I guess(ed) that he went* versus *I want(ed) him to go*). Gisborne (2009) argues that "if the finiteness contrast is levelled under verbs such as GUESS, then there is robust evidence that for (some) speakers of HKE, the morphosyntactic feature system of the verb is that of Cantonese, rather than that of the lexifier" (157). Likewise, a good indicator for identifying a tense contrast in HKE is the degree to which HKE verbs are used in their base form when they refer to the past and when they are not prone to conso-

[45] Unfortunately, no information on aspectual distinctions is given in that context, but simple aspect prevails in the examples the authors discuss.

[46] For an in-depth account of voices for and against a finiteness contrast in Chinese beyond the issue of tense marking, see Li (1986; 1990), Huang (1984; 1987; 1989; 1998), Hu et al. (2001), and Gisborne (2009: 157–159). The debate mainly evolves around the question whether Chinese mood and aspect distinctions "are realizations of finiteness or not" (Gisborne 2009: 157).

nant cluster reduction (ibid.: 160). In a short analysis of the verb *decide* in ICE-HK, Gisborne (2009: 161) identifies 63 hits of *decide* in a past tense context of which 47 instances are marked and 16 are not marked for past tense. The marked forms comprise tensed forms as well as participles; some of them occurring in passive constructions. Gisborne (2009) attributes this "considerable degree of variability" (166) in tense marking to the coexistence of and contact between English and Cantonese in Hong Kong. Gisborne (2009) observes different realizations of finiteness contrasts following verbs like *guess*, *suggest*, or *request* in ICE-HK (e.g., SUGGEST occurs both with finite (86 hits) and non-finite (twelve hits) complement clauses). Some of the realizations can be attributed to lexical distributions (e.g., non-finite complements occur with GUESS but not with REALIZE[47]) and register variation (e.g., many of the tokens of REALIZE occur in transcripts of legal cases). The lack of systematic finiteness contrasts is explained insofar as "English and Cantonese are still in contact in HK, and what we see is an emerging system with a considerable degree of variability" (ibid.: 166).

The question whether omission of past tense marking is phonologically conditioned is not discussed in the literature on HKE, but several authors mention final plosive deletion as a typical feature of HKE. The fact that syllable-final consonant clusters do not exist in Cantonese (cf. Matthews & Yip 1994: 19) often leads to pronunciation difficulties for native speakers of Cantonese. In their study on consonant cluster simplification in the speech of two native speakers of Cantonese from Hong Kong, Peng & Setter (2000) observe that alveolar plosives are often deleted in word-final consonant clusters, but the deletion rates differ considerably even among the two speakers investigated. The speech of each speaker is very consistent in its degree and patterns of consonant cluster reduction though. Unfortunately, no information is provided on the degree to which inflectional suffixes are affected by reduction. Bolton & Kwok (1990, in Hung 2000: 338) note that "final consonants are sometimes deleted." However, they neither provide details on the types of final consonants that are particularly affected, nor do they address deletion rates of inflectional suffixes. Examples for final consonant cluster reduction involving alveolar plosives are /ft/ that is reduced to /f/ and /kts/ that turns into /ks/ (compare also Bolton 2003: 208).

[47]It is not mentioned whether Gisborne (2009) searched for *realize*, *realise*, or both.

Indian English

As stated in section 2.1 and above, feature 132 (zero past tense forms of regular verbs) received a C-rating for IndE in eWAVE, meaning that the feature is estimated to occur extremely rarely. With that in mind, it is not surprising that lack of past tense marking has not received much attention in the literature on IndE. Exceptions are Sharma (2009) and Sharma & Deo (2009), who elaborate on past tense marking and lack thereof in perfective compared to imperfective contexts and on potential substratum influence underlying differences in marking.

The section on SgE above mentioned Sharma (2009: 176), who reports that SgE verbs in perfective contexts are marked for past tense to a larger extent than verbs in imperfective contexts (56.2 percent compared to 14.7 percent (habitual, progressive uses) and 36.9 percent (lexical stative uses), respectively). She observes an even more clear-cut distribution in her IndE data. 76.6 percent of the verbs in perfective contexts are marked for past tense compared with 29.5 percent in habitual, progressive contexts and 44.2 percent in lexical stative contexts. The data consist of sociolinguistic interviews with twelve non-English dominant Indians (ibid.: 174). All the Indians recorded are speakers of Hindi; two speakers additionally speak Gujarati and three Punjabi as another native language. Both substrate languages (Hindi in the case of IndE, Mandarin in the case of SgE) are aspect-prominent and use overt markers in perfective (completive) contexts; Mandarin uses the verbal particle *le*, Hindi uses *-(y)a*. Sharma (2009) explains the findings for both varieties as "straightforward instance[s] of strict transfer [...] whereby the semantic component of a form-meaning pairing in the L1 is re-attached to an L2 form" (179). She depicts this reattachment as follows:

(5.3) Hindi: [PERF *-a*] → IndE: [PERF *-ed*]

(5.4) Chinese: [PERF *le*] → SgE: [PERF *-ed*]

In contrast with Hindi and Chinese, (standard) English marks all verbs in a past tense context, irrespective of aspect (ibid.). A very important issue that Sharma (2009) additionally raises is that in contrast with SgE, where *already* is often used as a perfective marker, "IndE has not grammaticalized adverbs for aspectual functions" (179), which can also be explained by transfer. While Chinese and Malay use the analytic forms *le* and *sudah* as past tense markers, Hindi has no comparable analytic marker(s) for past tense.

As mentioned above for SgE, Silver et al. (2009) see the (Lexical) Aspect Hypothesis confirmed, according to which the (lexical) aspectual meaning of verbs impacts the degree to which those verbs are marked for tense and aspect. Sharma & Deo (2009) test the (Lexical) Aspect Hypothesis for IndE and note that learners of English with L1s that overtly mark perfectivity and imperfectivity (as Indo-Aryan languages do) are sensitive to sentence operators like time adverbials on top of verb meaning rather than to pure verb meaning in marking their verbs for past tense.[48] This confirms the authors' so-called Sentential Aspect Hypothesis, according to which "[l]earners hypothesize that morphological marking is a form of agreement with the aspectual class of the sentential predication (not narrowly with aspect alone)" (ibid.: 7). In contrast with Mandarin, which requires the marker *-zhe* in certain imperfective (non-progressive) contexts only, Hindi overtly marks imperfectivity (non-progressive, habitual) by means of *-ta* (cf. Sharma 2009). Consequently, speakers of Hindi constantly choose between the use of perfective and imperfective markers (Sharma & Deo 2009: 8). The resulting "sensitivity to perfectivity distinctions" (ibid.) makes it possible to test whether speakers of Hindi rely on "purely universal lexical aspect distinctions" (ibid.) or "retain a sensitivity to clausal (im)perfectivity when acquiring English as an L2" (ibid.).

Tickoo (2005) investigates the explanatory nature of the (Lexical) Aspect Hypothesis in her account of the selective marking for past tense in 35 narrative essays written by low intermediate-level learners of English from India. The learners are Hindi or Urdu mother tongue speakers and are described as "not skilled writers in any language" (ibid.: 373) who are "influenced by the more familiar practices of their oral language use" (ibid.). Tickoo (2005) points out that neither the (Lexical) Aspect Hypothesis nor the so-called grounding hypothesis can fully explain the observed patterns of selective marking. According to the grounding hypothesis, speakers use selective marking "to signal a distinction between two different types of narrative progression" (ibid.: 375), i.e., to differentiate between salient and less salient happenings (e.g., Kumpf 1984; Véronique 1987; von Sutterheim & Klein 1987).[49] An alternative explanation proposed is that selective marking "is likely to signal differences in modes of narration" (Tickoo 2005: 375) that are motivated by L1

[48] Compare Sharma & Deo (2009: 5–6) for a critical account of the (Lexical) Aspect Hypothesis.

[49] Hopper (1979) unravels these discourse phenomena by presenting a conglomerate of features typical of perfective (foregrounded, salient) and imperfective (backgrounded, less salient) contexts. Telicity, givenness, dynamism, saliency, and the realis/irrealies distinction are considered. For details see Hopper (1979: 216).

influence. Recorded Hindi oral narratives show that present tense is used for information that is "either given in the preceding discourse or inferable in the context of their occurrence" (ibid.: 372). It marks the "in-present-time experience of the past situation" (ibid.: 373). In fact, some speakers investigated tend to "reformulate[ ] [this] L1 narrative strategy of securing hearer engagement" (ibid.).

Consonant cluster reduction has been reported for IndE as well, although a potential effect on omission of past tense marking has not been explicitly mentioned to the author's knowledge. This is likely due to the fact that consonant clusters tend to be reduced more by speakers of Tibeto-Burman languages and less so by speakers of Indo-Aryan languages like Hindi. It is mainly speakers of Hindi, however, whose patterns of verbal past tense marking have been examined so far (see above). Wiltshire (2013) investigates word-final consonant devoicing and cluster reduction in five speakers of IndE with different L1 backgrounds (Hindi, Gujarati, Ao, Angami, and Mizo). Hindi and Gujarati are Indo-Aryan languages, whereas Ao, Angami, and Mizo belong to the Tibeto-Burman language family. She shows that variation among the speakers can be explained both by transfer of L1 phonotactics and by markedness constraints. Regarding L1 phonotactics, she observes that the Tibeto-Burman speakers devoice consonants and reduce consonant clusters far more often than the Indo-Aryan speakers (ibid.: 604–605). As to markedness, the Tibeto-Burman speakers tend to produce voiceless obstruents only (ibid.: 609–610), which are less marked than their voiced counterparts (e.g., Lombardi 1991, in Wiltshire 2014: 27). Particularly final consonant clusters that have less marked sonority sequencing are retained; i.e., clusters that fall in sonority towards the end of the syllable (e.g., Steriade 1982; 2001; Wiltshire 2013: 612–613).[50]

Khan (1991, in Wiltshire 2014) finds high consonant deletion rates among speakers of IndE from Aligarh in Uttar Pradesh (northern India), where Hindi is the main language spoken. He reports that consonant deletion occurs in particular before consonants (55.4 percent to 67.5 percent), pauses (20 percent to 24.5 percent), and vowels (10.2 percent to 20.4 percent; Wiltshire 2014: 32). The observed tendency

[50] Wiltshire (2013) accounts for markedness by means of Optimality Theory (OT; cf. Prince & Smolensky 1993) and The Emergence of the Unmarked (TETU; cf. McCarthy & Prince 1994). It suffices to say here that according to OT, "markedness constraints play a major role in determining an output, as they compete with constraints on correspondence to input or to related forms" (Wiltshire 2013: 599). When markedness constraints are high, a less marked output wins at the expense of correspondence to the input. TETU plays a role when correspondence constraints weigh higher than markedness constraints, but the correspondence is not active (e.g., in the case of epenthetic vowels for which no corresponding input exists; cf. ibid.: 599–600).

to delete clusters with mixed voicing (nt, lt, lk) is likely due to transfer (Khan 1991). According to Wiltshire (2014), the observation that speakers favor deletion when a fricative or sonorant (rather than a stop) precedes might be explicable by "a preference for deletion in bimorphemic forms" (ibid.: 33) in the sample because it runs counter to markedness assumptions.

5.2 Sample choice and hypotheses

Sample choice

As the previous section revealed, several factors that impact omission rates in the contact varieties of interest have been discussed in the literature, namely substratum transfer, the phonetic environment, verb meaning, and perfective versus imperfective verb use. As to the phonetic environment, we have learned that consonant cluster reduction is likely a cause of lack of past tense marking in SgE and HKE; as well as among Tibeto-Burman speakers of IndE. Regular verbs for which the inflectional suffix is the sole past tense marker are affected, irregular verbs that form their past tense by means of vowel change and the addition of [t,d] not. Since the original ICE recordings are not available, the analyses here focus on regular verbs that end in a vowel to exclude any impact of the phonetic environment on omission rates. Additionally, cases in which the regular verb of interest is followed by a consonant-initial word were excluded from the analyses. Recall examples 5.1 and 5.2 in the introductory lines to this chapter, where the regular *-ed* suffix would occur between vowels or semivowels (*show uh* and *stay in*, respectively). Nevertheless, in a first step verbs whose lack of the past tense /t,d/ suffix can be phonologically conditioned were accounted for in order to get a rough idea of omission rates in consonant-final compared to vowel-final and irregular verbs. For the reasons just mentioned, the respective results need to be interpreted with caution.

Due to the focus on regular verbs ending in vowels, samples of regular (and irregular) verbs in the spoken part of ICE and in GloWbE were chosen instead of investigating subsets of ICE spoken and GloWbE for all regular (and irregular) verbs occurring there. Working with subsets of GloWbE would not have been feasible in any case. The approach taken did not only increase the token numbers to work with but also made it possible to account for register differences. Additionally, speaker bias was reduced.

A major challenge of investigating inflectional marking is that corpus searches need to be conducted on open-class items, which means that comprehensive automatic searches result in considerable amounts of redundant data (compare Ziegeler 2015: 184 for SgE). When those open-class items lack inflectional marking, therefore constituting bare forms, automatic searches become impossible. This is why all marked and unmarked verbs were retrieved manually.

In order to take verbs of different frequencies of use into account, lists of the lemma token frequencies of all verbs in GloWbE were extracted from the GloWbE full-text offline database for each of the varieties of interest (compare section 4.4.1). Given the large size of GloWbE, this approach allowed for a more detailed analysis of lemma token frequencies than if only frequency lists derived from ICE had been considered. A comparison of the extracted frequencies in GloWbE with those in ICE revealed that verbs below a frequency of occurrence of around 1,000 in GloWbE are unlikely to occur in ICE at all. This is why verbs below this frequency threshold in GloWbE were discarded right away. 200 vowel- and consonant-final regular and 75 irregular verbs of varying token frequencies remained, from which random samples of 20 irregular, 20 vowel-final, and 20 consonant-final regular verbs were drawn. This left regular and irregular verbs of varying token frequency. A forced binary distinction between frequent and infrequent verbs was explicitly decided against because the verbs' frequency distribution made it difficult to draw the line somewhere meaningful.[51]

As section 5.1 revealed, substratum transfer from Chinese dialects and Indo-Aryan languages influences patterns of past tense omission in perfective and imperfective contexts. This aspect as well as verb meaning were briefly accounted for in the analyses. The focus on 20 irregular and 20 vowel-final regular verbs (plus 20 consonant-final regular verbs that were considered separately), and the priority to sample verbs along the frequency cline, made it impossible to include verb meaning as another sampling factor though.

The sampled vowel-final regular verbs are *play*, *agree*, *allow*, *stay*, *show*, *follow*, *apply*, *enjoy*, *carry*, *die*, *identify*, *continue*, *view*, *issue*, *argue*, *destroy*, *deny*, *employ*, *reply*, and *rely*. The sampled irregular verbs are *think*, *know*, *go*, *see*, *mean*, *take*,

[51] Initially, the 200 regular and 75 irregular verbs had been divided into quartiles. Verbs in the quartile with the lowest token frequencies proved to be too infrequent to work with though, and the verb frequencies in the remaining quartiles made it difficult to distinguish between frequent and infrequent verb lemmata. This shows the challenges of working with a relatively small corpus and makes clear why it was necessary to consider all spoken sections in ICE for the analyses.

come, *tell*, *stick*, *catch*, *seek*, *throw*, *fight*, *fall*, *wear*, *grow*, *break*, *drive*, *forget*, and *begin*. The sampled consonant-final regular verbs are *call*, *live*, *happen*, *help*, *watch*, *join*, *finish*, *sign*, *maintain*, *establish*, *claim*, *pull*, *stop*, *develop*, *expect*, *want*, *visit*, *need*, *like*, and *wish*. *Develop*, *expect*, *want*, and *need* are special cases because they form their past tense by means of [ɪd] affixation, i.e., by adding an extra syllable.

The English verb phrase is a complex construct. In contrast with non-finite verb phrases, which are not specified for tense and modality (cf. Biber et al. 2007: 99), finite verb phrases are structurally distinguished on the basis of tense (present, past), aspect (simple, perfect, progressive, perfect progressive), voice (active, passive), modality (tensed, modal), negation (positive, negative), and clause structure type (declarative, interrogative; cf. Biber et al. 2007: 452). Past is only one of the formal features that marks the functional category of tense, the other two being agreement and finiteness (e.g., Leung 2003: 200). For speakers with a Chinese mother tongue background, this is a major difficulty because the lack of tense and agreement as grammatical categories in Chinese makes it necessary to acquire both those functional categories and their formal features. In Malay, overt tense or agreement morphology is lacking as well. The Indian substrate languages of interest have verbal past tense marking but they are aspect- rather than tense-prominent like Chinese.

In English, inflectional past tense marking occurs both in finite (e.g., *He played*) and in non-finite verbs (*He has played*, *He had played*, *The ball is (being) played*, *The ball was (being) played*, *The ball has been (being) played*, *The ball had been (being) played*). Crucially, only finite verbs specify for tense though. The corpus study presented here makes a distinction between omission of inflectional past tense marking in finite and non-finite uses of the sampled verbs to account for past tense marking both formally and functionally. Concerning the latter, a major point of interest is the question whether past tense omission is particularly likely when the respective verb is accompanied by a time adverbial with past time reference. All functional uses of verbal past tense marking constitute formal uses as well, but not vice versa. Example 5.5 below is a case of formal past tense marking, example 5.6 one of functional past tense marking including a time adverbial with past time reference.

(5.5) Oh I would have *take* him (ICE-IND:S1A-019#139:1:B)

(5.6) Yeah last year we also *apply* our group also applied for touch camp leader (ICE-HK:S1A-005#305:1:B)

A past tense context was defined in line with Gut (2009b: 265–266), namely either when a time adverbial is present or when the past time reference is obvious from the surrounding context, e.g., when a past time activity is being reported. Cases where markup (e.g., extra-corpus material (<X></X>), quotations (<quote></quote>), or uncertain transcriptions (<?></?>)) directly affected the forms of interest were not counted. Additionally, wrongly marked forms and self-corrections were ignored.[52] Also, instances of narrative present were excluded from the analyses—irrespective of whether the resulting present tense form is finite or non-finite (distinguishable for third person singular only). Lastly, it was made sure that in none of the ICE corpora considered past tense omission was restricted to a certain number of speakers.

Hypotheses

As pointed out before, the impact of three factors on past tense omission is of particular interest here, namely frequency of use, substratum transfer, and institutionalization. Let us reconsider the respective underlying hypotheses introduced in section 1.3.

As discussed in section 2.2, usage-based approaches to language deal with the effects frequencies of use have on language acquisition, use, and change. In that vein, the underlying assumption was that frequencies of use impact degrees of past tense omission. As frequency measures, both lemma token and type frequencies were considered.[53] Those frequency measures were approximated for each variety by means of the abovementioned frequency lists extracted for the sampling purposes from the GloWbE full-text offline database. The respective ICE frequencies were not considered to guarantee an independent frequency database that does not intermingle with the analyses conducted in ICE.[54] While the lemma token frequency for each verb was directly retrieved from the respective GloWbE list, its type frequency was determined the following way: Each lemma was assigned a group according to the morphological process involved in forming its past tense ([t,d] affixation, [ɪd] affixation, suppletion, vowel change, vowel change + [t,d]; cf. Gut 2009b: 267). The number of lemmata per group defines the type frequency. Since GloWbE HK, GloWbE SG, and GloWbE IN differ in size, relative frequencies were used (frequency by corpus size).

[52] I.e., in an utterance such as *What happen what has happened in the past to these particular family okay* (ICE-HK:S1B-010#24:1:A) the unmarked form *happen* was counted. The corrected marked form *has happened* was not counted as an instance of formal past tense marking instead.

[53] For definitions of both frequency types, see section 2.2.

[54] Many thanks to Susanne Wagner (p.c., 20 September 2016) for valuable feedback on this matter.

Since regular as well as irregular verbs were accounted for, both token and type frequency effects could be examined. Irregular verbs tend to occur more frequently than regular verbs, but they are of relatively low type frequency. As elaborated on in section 2.2, two seemingly contradictory frequency effects have been reported for language change in the literature. On the one hand, highly frequent forms are less prone to change than infrequent forms due to their strong entrenchment in the mind (Conserving Effect; e.g., Bybee 2007: 10). On the other hand, highly frequent forms are prone to reduction first, infrequent forms only later (Reduction Effect; e.g., Bybee 2007: 11). For language acquisition, it has been observed that frequent forms are acquired first, infrequent forms later. Due to the focus on verbs whose lack of past tense is not phonologically conditioned, only the Conserving Effect resulting from previous observations that frequent forms are acquired earlier and are more strongly entrenched than infrequent forms is of importance here. Putting aside the stronger entrenchment of irregular compared to regular verbs for a moment, the reasoning implies that frequent regular verbs are prone to omission less (or later) than infrequent regular verbs. Depending on the question whether omission is a learner feature or a variety-specific innovation (Van Rooy 2011: 192; see section 2.4), past tense omission was expected to affect infrequent verbs more (learner feature) or first (variety-specific innovation) and frequent verbs less or later. This is the underlying hypothesis (compare section 1.3):

Hypothesis 1a *In cases where omission of inflectional marking is morphologically conditioned, infrequent forms are affected by omission more (or first) and frequent forms less (or later). This is the case for both regular and irregular forms.*

The choice of varieties allowed accounting for the impact of both substratum transfer and degree of institutionalization on omission rates. Turning to substratum transfer first, we have learned that Chinese dialects and Hindi are aspect-prominent. Sharma (2009) observes considerably lower omission rates in perfective contexts than in imperfective contexts, which she explains with direct transfer of verbal particles that indicate perfectivity in Mandarin and Hindi to English past tense marking. Bao (1995) discusses the perfective marker *already* in SgE, which he traces back to Chinese *le* and Malay *sudah*. For HKE, Wong (2017: 16, with reference to Matthews & Yip 1994: 198) points out that the Cantonese way to mark time by adverbials, among other things, is transferred to the contact variety. In contrast with the Indian substrate languages of interest (compare section 5.5), Chinese additionally lacks verbal

tense marking, which means that speakers of English with L1 Chinese background need to acquire tense as a grammatical category. Since Chinese dialects "share a common core in grammar and vocabulary" (Chao 1968, in Bao 2010: 794), SgE and HKE speakers are equally affected. Consonant cluster reduction as a result of lack of word-final consonant clusters in Mandarin, Cantonese, and Tibeto-Burman languages can account for past tense omission as well, but only verbs that end in a consonant other than an alveolar plosive are affected. It was assumed that in case SgE and HKE are more strongly affected by verbal past tense omission than IndE, influence from common isolating substrata (that are not only aspect prominent but actually lack verbal past tense marking) accounts for the observed omission rates. Substratum transfer was expected to constrain frequency effects in case omission occurs considerably more in SgE and HKE than in IndE irrespective of lemma token frequency.

Hypothesis 2 *Substratum transfer functions as a constraint on frequency effects in case omission and regularization occur considerably more in SgE and HKE than in IndE irrespective of lemma token frequency, given that the observed simplification patterns can be explained by common substrate influence.*

As to institutionalization, the degree of stability of the contact varieties is of particular interest here. Coming back to Van Rooy's (2011) account of learner errors versus conventionalized innovations (section 2.4), social factors like acceptance are of crucial importance for the emergence of conventionalized innovations. Consequently, we can deduce that past tense omission only constitutes a conventionalized innovation in case it is (grammatically) stable and accepted. The former was tested by means of the available corpus data, the latter by means of the perception experiment described in chapter 8. If there is a tendency for past tense omission, the feature was expected to be most stable in SgE, the most institutionalized variety accounted for here. SgE is on its way to stage 5 in Schneider's (2003; 2007) Dynamic Model in contrast with HKE (stage 3) and IndE (stage 3, arguably stage 4). Hypothesis 3 was formulated as follows (compare section 1.3):

Hypothesis 3 *Institutionalization functions as a constraint of frequency effects in case omission and regularization patterns are particularly stable in SgE irrespective of lemma token frequency.*

5.3 Omission rates: An overview

Omission rates by corpus and morphological process

Let us start with an overview of degrees of past tense omission in regular and irregular verbs in the spoken part of ICE.[55] As discussed above, the lack of availability of the original ICE recordings made it necessary to focus on vowel-final regular verbs followed by a vowel-initial word or a pause. That way, past tense omission as a result of consonant cluster reduction was ruled out. Irregular verbs keep their past time reference irrespective of the sound the next word starts with. Besides vowel-final regular verbs and irregular verbs, verbs that end in an alveolar plosive are further candidates whose past tense formation is not affected by consonant cluster reduction, but the fact that an extra syllable is added to mark them for past tense makes them a special case. This is why they were not considered either.

Figure 5.1 depicts the distribution of the omission rates by morphological process and corpus. All types of morphological past tense markings are displayed to get a rough idea of the degree to which consonant cluster reduction accounts for past tense omission. Box plots are a means of visualizing data distributions. The boxes depict the middle 50 percent of the data, or in other words, the interquartile range that contains data points from the first to the third quartile (cf. Baayen 2012: 30). The horizontal line within the box represents the median value. While the whiskers maximally comprise "1.5 times the interquartile range" (ibid.), all points outside the whiskers are clear outliers.

The omission rates were calculated by dividing the number of verbs not inflectionally marked for past tense by the sum of verbs that are marked as well as not marked for past tense. Table B.1 in appendix B provides the number of verbs marked and not marked for past tense in ICE-SIN, ICE-HK, and ICE-IND by usage type and morphological process. Table B.2 (appendix B) lists for each sampled verb the number of marked and unmarked forms as well as the resulting omission rate and the lemma token frequency in ICE; for both formal and functional uses. Regarding vowel- and consonant-final verbs with [t,d] affixation, figure 5.1 makes a distinction between verbs that are followed by a vowel-initial word (or a pause) and verbs that are followed by a consonant-initial word. Additionally, based on the assumption that consonant cluster reduction affects past tense marking formally (i.e., irrespective of

[55] Henceforth, whenever the ICE corpora are referred to, only the spoken sections are meant.

verb function), all cases of missing verbal past tense marking were considered; including both simple past forms and past participles.

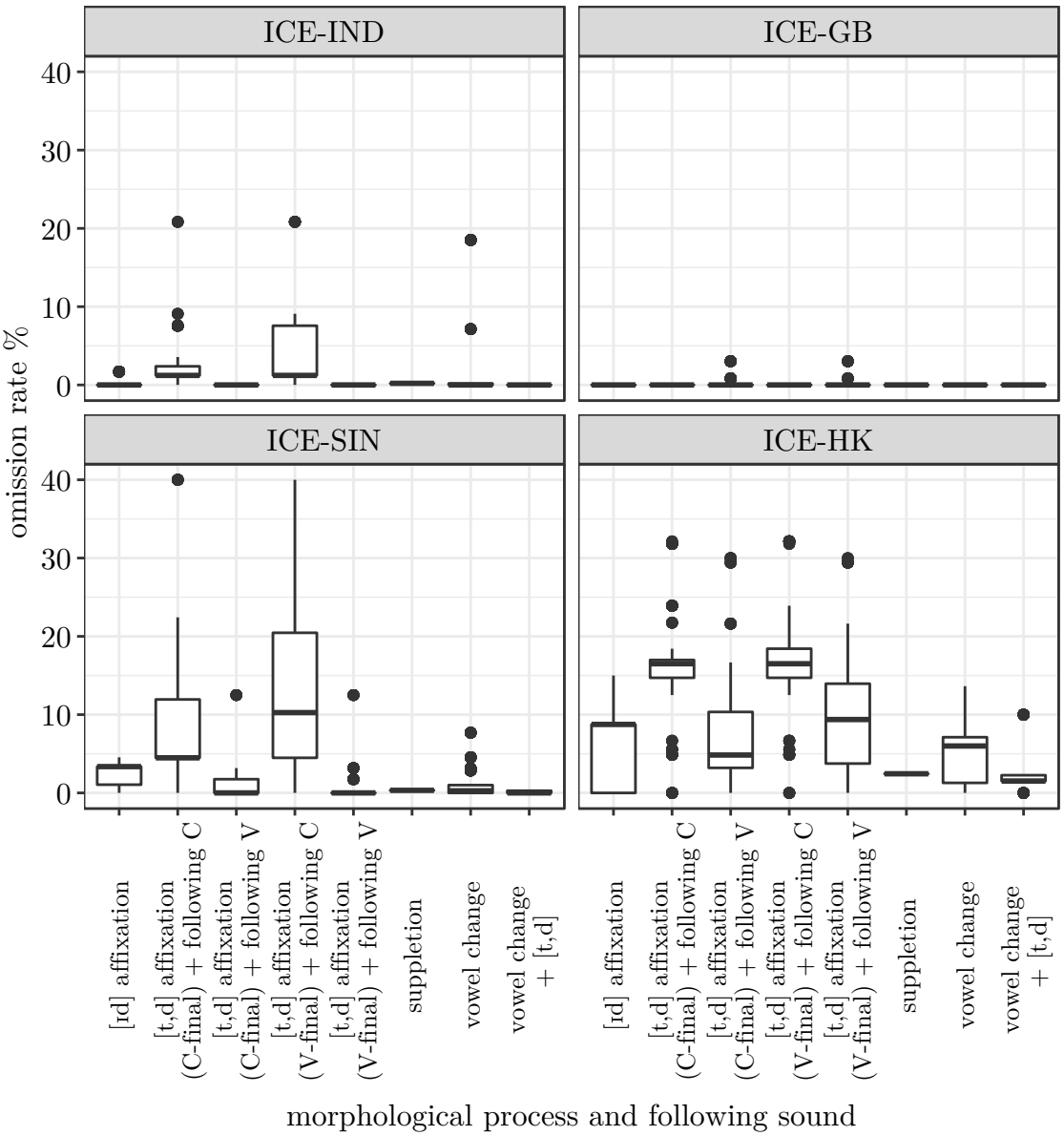

Figure 5.1: Past tense omission rates by morphological process and following sound, by corpus (formal uses)

We observe comparatively high omission rates in ICE-HK and ICE SIN (only among the regular verbs in the latter case[56]) and small omission rates in the Indian

[56]Recall that Gut (2009b: 275) observes more omission of single-final and cluster-final /t,d/ in other lexical verbs than in the *-ed* morpheme in SgE. She concludes that "highly educated speakers of SgE" (275) are able to actively suppress /t,d/ deletion when the past tense suffix

component of ICE and in ICE-GB. However, with the exception of a few outliers omission rates hardly exceed 20 percent across corpora. Except for ICE-GB, where omission is basically non-existent, omission rates are highest when a consonant-initial word follows. In ICE-HK, both consonant-final and vowel-final verbs followed by a consonant-initial word have considerably higher omission rates than consonant-final and vowel-final verbs followed by a vowel-initial words (or a pause) and verbs with [ɪd] affixation. This is a sign of consonant cluster reduction despite the fact that consonant-final verbs followed by a vowel-initial word (or a pause) have surprisingly low omission rates. This would be worth checking were the recordings available. The omission rates that define the picture for vowel change verbs can be attributed to the lemmata *begin*, *come*, *fight*, *forget*, *stick*, *take*, and *throw*.[57] In contrast with omission in ICE-HK, omission in ICE-SIN is nearly exclusively restricted to CCR environments. The comparatively high omission rates for vowel-final verbs with [t,d] affixation followed by a consonant-initial word in ICE-IND are mainly due to omission of the inflectional suffix in *happen* (omission rate: 7.57, 171 marked forms, 14 unmarked forms). They are not speaker-specific.

The overall picture clearly indicates that in ICE-HK and ICE-SIN consonant cluster reduction contributes to lack of verbal past tense marking. While the observed omission rates in the direct proximity of consonants have to be interpreted with caution, the fact that relatively many verbs were transcribed as unmarked for past tense in such environments is telling. In all three target varieties, the many outliers are signs of the large variation in omission rates across lemmata.

Syllable number

An additional aspect tested was the effect of syllable number on omission rates. For each lemma, the number of phonetic syllables was adopted from WebCelex (Max Planck Institute for Psycholinguistics 2001). However, it has to be pointed out that the lack of availability of the ICE transcripts made it impossible to check whether the WebCelex syllable counts match the number of syllables in the original ICE recordings. For exactly this reason, syllable number was no criterion in the sampling process. Only the vowel-final regular verbs followed by a vowel-initial word were

is affected. Although it is not explicitly stated, the examples provided by Gut (2009b) indicate that formal rather than functional uses of the past tense suffix are meant.

[57] The omission rates (marked for past tense: not marked for past tense) are as follows: *begin* 7.69 (36:3), *come* 7.12 (248:19), *fight* 10.00 (9:1), *forget* 13.64 (38:6), *stick* 11.11 (8:1), *take* 5.99 (345:22), and *throw* 6.67 (14:1). I.e., the verb *begin* has an omission rate of 7.69 percent, with 36 marked and three unmarked forms with past time reference, etc. This notation will be used henceforth.

considered here because the irregular verbs are a too heterogeneous class to account for the impact of syllable number on omission. Consonant-final verbs with /t,d/ affixation, verbs with /t,d/ affixation followed by a consonant-initial word, and verbs with [ɪd] affixation were not taken into account for the reasons mentioned above. Table 5.1 depicts the observed omission rates by corpus and syllable number.

Table 5.1: Past tense omission rates by corpus and syllable number (formal uses, vowel-final regular verbs + following V only)

corpus	omission rate % (marked:not marked)							
	1 syllable		2 syllables		3 syllables		4 syllables	
ICE-SIN	0.00	(83:0)	1.54	(128:2)	0.00	(7:0)	0.00	(8:0)
ICE-HK	24.19	(47:15)	5.56	(119:7)	0.00	(5:0)	0.00	(11:0)
ICE-IND	0.00	(121:0)	0.00	(181:0)	0.00	(11:0)	0.00	(25:0)

Unsurprisingly, most marked and unmarked forms are mono- or bisyllabic because the majority of the sampled consonant-final regular verbs fall in that category. Omission rates are comparatively high for monosyllabic verbs in ICE-HK. The respective unmarked lemmata are *allow*, *die*, *play*, *show*, and *stay*, unmarked *play* occurring twice in the same utterance and *stay* being uttered three times by the same speaker. Whether the observation that monosyllabic verbs in ICE-HK are particularly prone to omission is more than a chance observation deserves further attention but cannot be investigated here. A larger sample of regular verbs and access to the original recordings are necessary for that.

Formal versus functional uses of verbal past tense marking

Figure 5.2 focuses on ICE-HK and contrasts omission rates in all verbs that should be formally marked for past tense with omission rates in verbs that should be realized as simple past forms (functional marking) exclusively. As mentioned before, the latter are a subgroup of the former. Lack of verbal past tense marking in functional uses in ICE-SIN and ICE-IND proved to be too small to work with. Again, consonant-final verbs with /t,d/ affixation, verbs with /t,d/ affixation followed by a consonant-initial word, and verbs with [ɪd] affixation were not considered. Regarding /t,d/ affixation verbs followed by a vowel-initial word and vowel change verbs, omission rates are higher when functional uses of verbal past tense marking are considered exclusively than when all formal uses are taken into account. An interesting question in that context is whether the presence of time adverbials in the immediate surroundings

promotes omission of verbal past tense marking in the sense that an accompanying time adverbial accounts for the past time reference and makes verbal past tense marking redundant. For previous research on that issue see section 5.1.

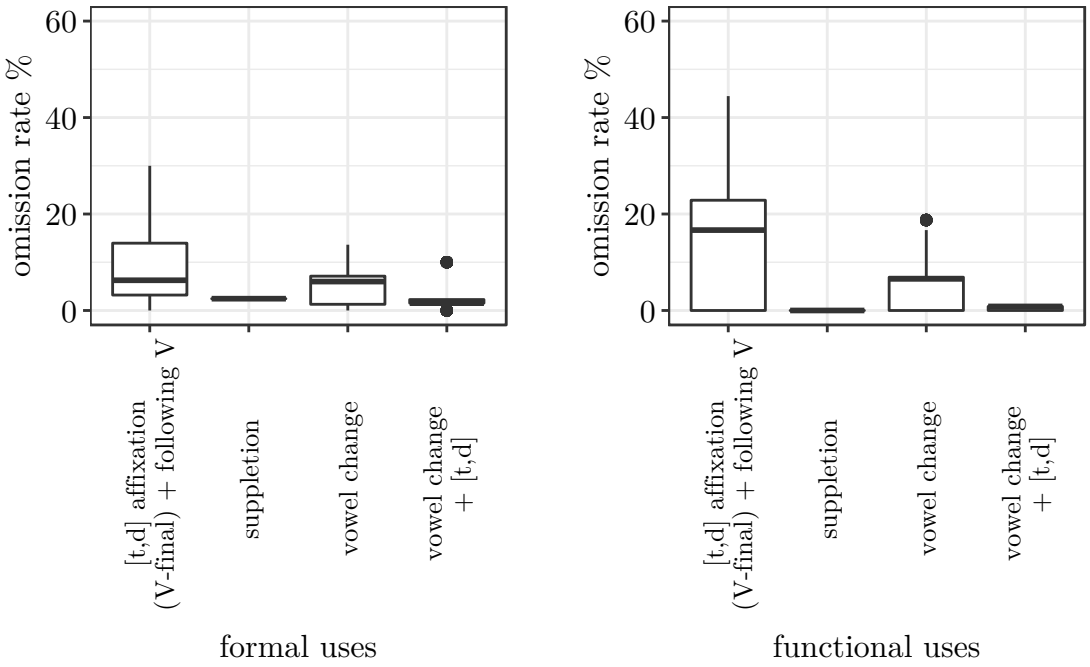

Figure 5.2: Past tense omission rates by morphological process and following sound (subset) in ICE-HK, by usage type

Table 5.2 summarizes the number of unmarked verbs preceded, followed, or not accompanied by a time adverbial in ICE-HK. As figure 5.2 revealed, the absolute number of unmarked verbs that lack functional past tense marking is much smaller than that of verbs that lack formal past tense marking. While a considerable number of the sampled verbs that lack functional past tense marking is preceded or followed by a time adverbial, it is impossible to estimate the impact of time adverbials on past tense omission on the basis of the small number of unmarked forms observed. Even more, while corpus data allow detecting tendencies, it is only by means of controlled conditions that those tendencies can be investigated thoroughly. The experiment described in chapter 8 tested the impact of preceding time adverbials on the perception and judgment of lack of verbal past tense marking. Its focus is on preceding time adverbials to account for the on-line perception of unmarked forms.

Time adverbials present different time-related meanings (cf. Biber et al. 2007: 777). They can describe a position in time (e.g., *yesterday*, *last July*), refer to the duration of an event (e.g., *for years*), describe the temporal relationship between

Table 5.2: Number of verbs lacking functional past tense marking in ICE-HK, by morphological process and time adverbial

morphological process	time adverbial		
	preceding	following	no
[t,d] affixation (V-final, + following V)	7	2	9
vowel change	16	9	16
vowel change plus [t,d]	0	0	0

events or states (e.g., *before this*, *after that*) or provide information on the frequency with which an event occurs (e.g., *occasionally*, *very often*). Obviously, adverbials in the latter category (frequency of an event) do not indicate past time reference, which is why they were not of interest here. In example 5.6 (section 5.2), the inflectionally unmarked verb *apply* was preceded by the time adverbial *last year* (*Yeah last year we also apply our group also applied for touch camp leader*). In example 5.7 below, the unmarked verb is preceded by *after*, which accounts for the temporal relationship expressed. In 5.8, the time adverbial follows rather than precedes the target verb. In 5.9, the adverbial *never* expresses the duration of the event referred to.

(5.7) What we did was uh after installation of the EPCON system we *allow* it the system to be run in the factory for a few days (ICE-HK:S2B-041#84:1:A)

(5.8) It all *begin* with the uh < ? > spare </?> technology <.> o </.> after the second world war (ICE-HK:S2A-054#5:1:A)

(5.9) I never *think* that he's <unclear> word </unclear> until I <unclear> word </unclear> him in performing ah (ICE-SIN:S1A-065#171:2:F)

Let us have a look at omission of verbal past tense marking in GloWbE next. Only lack of functional past tense marking was focused on because the size of GloWbE makes it necessary to investigate neat categories. Considering lack of formal marking for past tense would have meant including participles occurring in passive constructions, which are not of interest here (e.g., *The ball is (being) played*; compare section 5.2). For each of the 60 sampled verbs and each variety (section 5.2), random samples of bare verb forms that potentially lack inflectional past tense marking were drawn by means of the search mask available in the online version of GloWbE. Bare forms of the lemma *call*, for instance, were searched by means of the following syntax: "call.[v*]". A click on the number of hits per variety (or per country, for that

matter; see section 4.1) listed all hits (KWIC display). By means of the sampling function incorporated in the web-based interface, samples of 200 hits per verb were drawn from GloWbE SG and GloWbE HK and samples of 400 hits from GloWbE IN. GloWbE IN is about twice the size of GloWbE SG and GloWbE HK. Initial searches had revealed no differences in omission rates between the general and the blog sections, which is why both sections were accounted for (see section 4.1). From the number of hits of verbs that lack inflectional past tense marking in each sample, the respective number of hits in the whole corpus was extrapolated. For instance, for *finish*, 11 unmarked forms were identified in the sampled 200 bare forms in GloWbE HK. This makes 99.94 unmarked forms among all 1,817 bare forms of *finish* in the corpus.[58] Finally, the omission rate (8.81) was calculated by dividing the extrapolated number of unmarked forms (99.94) by the sum of this number and the number of inflectionally marked simple past forms of *finish*, here 1,035 ("finished.[vvd]").[59] Figure 5.3 depicts the omission rates obtained for all the 60 lemmata of interest by corpus and verb type (regular verb ending in a vowel, regular verb ending in a consonant, irregular verb).

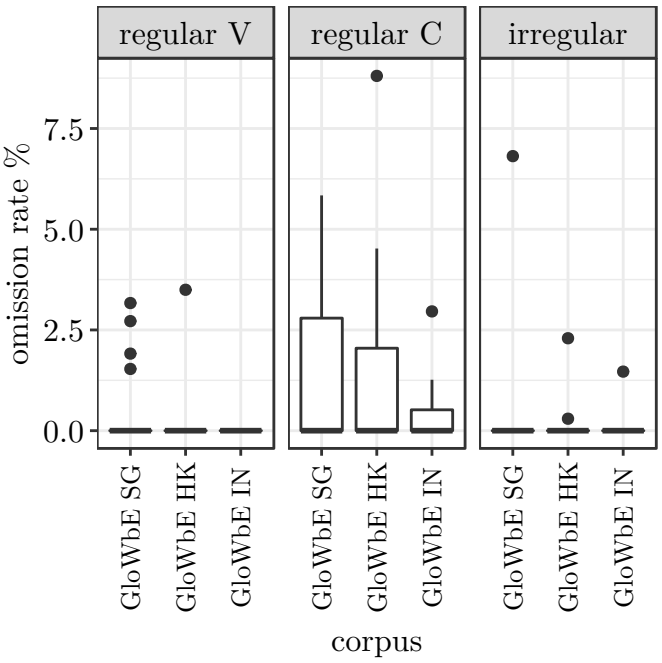

Figure 5.3: Omission rates in GloWbE (by verb type, functional uses only)

[58]The product of 11 unmarked forms and 1,817 bare forms was divided by the sample size of 200.
[59]Table B.3 in appendix B provides the respective numbers for all sampled lemmata.

Omission rates in GloWbE are very small across varieties and verb types. For most lemmata, the sampled bare verb forms contain no verbs that lack inflectional past tense marking. The findings clearly show that omission of verbal past tense marking is very much a feature of spoken language that has not found its way into internet language. Nevertheless, it is noteworthy that consonant-final regular verbs are comparatively much affected by omission in GloWbE; as they are in ICE. A word of caution is necessary here though. To account for all 60 lemmata, which was necessary in order to compare the GloWbE findings with those in ICE, small sample sizes had to be chosen. This means that single inflectionally unmarked verbs in those samples weigh heavily. Additionally, sample sizes of 200 and 400 were used irrespective of the overall number of bare forms of a lemma in GloWbE. However, going through the sampled bare forms manually left the impression that the large majority constitute standard bare forms. Since they were not of interest for this study, those bare forms were not classified further.

Genre-(un)specificity

Let us have a look at the distribution of the obtained omission rates across the different ICE sections (table 5.3 below). Only instances of functional past tense marking and lack thereof were considered. Recall from section 4.1 that all spoken sections in ICE were taken into account in the analyses. The distribution of the sections defines their relative size. Private dialogues, for instance, constitute 33.33 percent of the spoken part of ICE.

In ICE-HK, omission of verbal past tense marking prevails in face-to-face conversations, the most informal corpus section. 54.55 percent of all instances of lack of verb past tense marking in the verb samples accounted for were found there, followed by public dialogues and scripted monologues. The large number of unmarked irregular verbs in face-to-face conversations collected for ICE-HK is particularly remarkable. In ICE-SIN and ICE-HK omission rates are too low to account for any section-specific patterns. The findings support the decision to consider all spoken sections in ICE.

5.4 A usage-based approach to omission of verbal past tense marking

Let us turn to a usage-based account of omission of verbal past tense marking next. Since past tense omission in consonant-final regular verbs and in vowel-final regular

Table 5.3: Omission rates of vowel-final regular verbs and irregular verbs by corpus section (functional uses only)

corpus section	omission rates % (marked:not marked)				distribution of sections %
	vowel-final regular + following V		irregular		
ICE-SIN:					
Private dialogues	4.00	(48:2)	0.35	(1,155:4)	33.33
> face-to-face conv.	4.55	(42:2)	0.30	(1,012:3)	
> phonecalls	0.00	(6:0)	0.69	(143:1)	
Public dialogues	0.00	(16:0)	0.00	(217:0)	26.67
Unscripted monologues	0.00	(17:0)	0.39	(255:1)	23.33
Scripted monologues	0.00	(16:0)	0.00	(221:0)	16.67
ICE-HK:					
Private dialogues	48.00	(13:12)	5.52	(548:32)	33.33
> face-to-face conv.	54.55	(10:12)	6.04	(451:29)	
> phonecalls	0.00	(3:0)	3.00	(97:3)	
Public dialogues	21.05	(15:4)	2.94	(264:8)	26.67
Unscripted monologues	0.00	(21:0)	0.26	(380:1)	23.33
Scripted monologues	12.50	(14:2)	1.05	(189:2)	16.67
ICE-IND:					
Private dialogues	0.00	(37:0)	0.61	(649:4)	33.33
> face-to-face conv.	0.00	(34:0)	0.54	(550:3)	
> phonecalls	0.00	(3:0)	1.00	(99:1)	
Public dialogues	0.00	(30:0)	0.00	(361:0)	26.67
Unscripted monologues	0.00	(27:0)	0.62	(322:2)	23.33
Scripted monologues	0.00	(32:0)	0.00	(172:0)	16.67

verbs followed by a consonant-initial word is likely phonologically conditioned, we will concentrate on vowel-final regular verbs followed by a vowel-initial word as well as on irregular verbs henceforth again. The question of interest is whether past tense omission can be explained by usage frequencies, i.e., whether the fact that a verb is frequent or infrequent determines the degree to which it is affected by omission. For each variety, frequencies of use were approximated by the relative lemma token frequencies in the respective GloWbE corpus. Figure 5.4 depicts the omission rates by logarithmically transformed relative lemma token frequency for ICE-SIN, ICE-HK, and ICE-IND.

Formal uses of verbal past tense marking are accounted for in figure 5.4; vowel-final regular verbs preceded by a vowel-initial word or a pause and irregular verbs are depicted separately. The logarithmic transformation of the relative lemma token frequencies helps visualize the data because it makes the data points spread evenly across the graph rather than being tightly clustered in frequency regions many lemmata fall into. Each dot depicts a lemma.

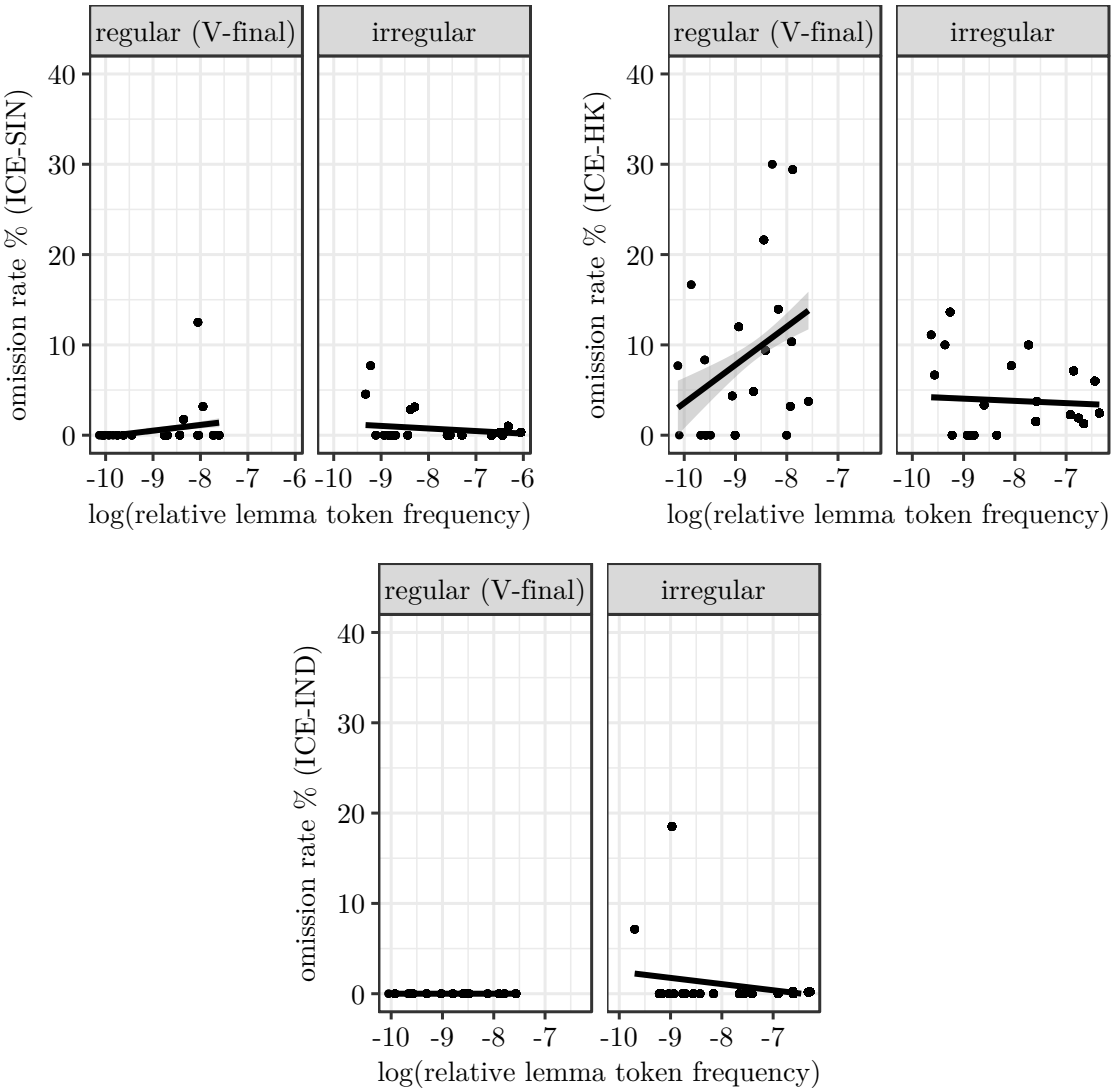

Figure 5.4: Past tense omission rates in ICE by relative lemma token frequency, by corpus (formal uses)

While in ICE-IND no omission is observable in vowel-final regular verbs, in ICE-SIN and ICE-HK regular verbs of higher relative lemma token frequency tend to have higher omission rates. In ICE-SIN, one clear outlier (*stay*; omission rate: 12.5,

marked: 14, not marked: 2) accounts for that trend. The opposite trend is observable for irregular verbs, but in ICE-SIN and ICE-IND the tendency is carried by a few outliers (*stick*, *throw*, *grow*, and *begin* in ICE-SIN; *stick* and *forget* in ICE-IND). In ICE-HK, the sampled regular verbs show large dispersion in their omission rates irrespective of their frequency, whereas the omission rates of the sampled irregular verbs are less scattered among the more frequent than among the less frequent verbs.[60] The regression lines in figure 5.4 visualize those trends.

The log-transformation of the relative lemma token frequencies distracts from the fact that across corpora the irregular verbs occupy a much larger frequency range than the regular verbs. Table 5.4 provides the omission rates for regular and irregular verbs in ICE-SIN and ICE-HK below a frequency threshold (relative frequency) of 0.00075, which the regular verbs in both corpora do not pass.

Table 5.4: Omission rates of regular and irregular verbs of relative lemma token frequency below 0.00075 in GloWbE (formal uses)

corpus	omission rate % (marked:not marked)			
	vowel-final regular + following V		irregular	
ICE-SIN	0.88	(226:2)	0.57	(692:4)
ICE-HK	10.78	(182:22)	2.98	(683:21)

In GloWbE SG, the frequency threshold of 0.00075 is equivalent to an absolute frequency below 32,231 and in GloWbE HK to an absolute frequency below 30,338. Omission in verbs below this frequency threshold is considerably more frequent in ICE-HK than in ICE-SIN, and in both corpora omission rates in vowel-final regular verbs surpass those in irregular verbs. Note how many irregular verbs are not marked for past tense in ICE-HK.

Figure 5.5 shows the omission rates by past tense rate (rather than by relative lemma token frequency) for ICE-SIN, ICE-HK, and ICE-IND. The past tense rate of a lemma was calculated by dividing the number of occurrences of the lemma in a past tense context (marked or unmarked) by the overall number of occurrences of the lemma in ICE (compare table B.2 in appendix B). Thus, the focus did not lie on the general frequency of occurrence of that lemma but on its frequency of occurrence

[60] The omission rates (marked for past tense: not marked for past tense) are as follows: for ICE-SIN, *stick* 4.55 (21:1), *throw* 7.69 (12:1), *grow* 3.13 (31:1), and *begin* 2.86 (34:1); for ICE-IN, *stick* 7.14 (13:1) and *forget* 18.52 (22:5) (compare table B.2 in appendix B).

in a past tense context. This is why functional uses of verbal past tense marking (i.e., marking for simple past and lack thereof) were taken into account exclusively.

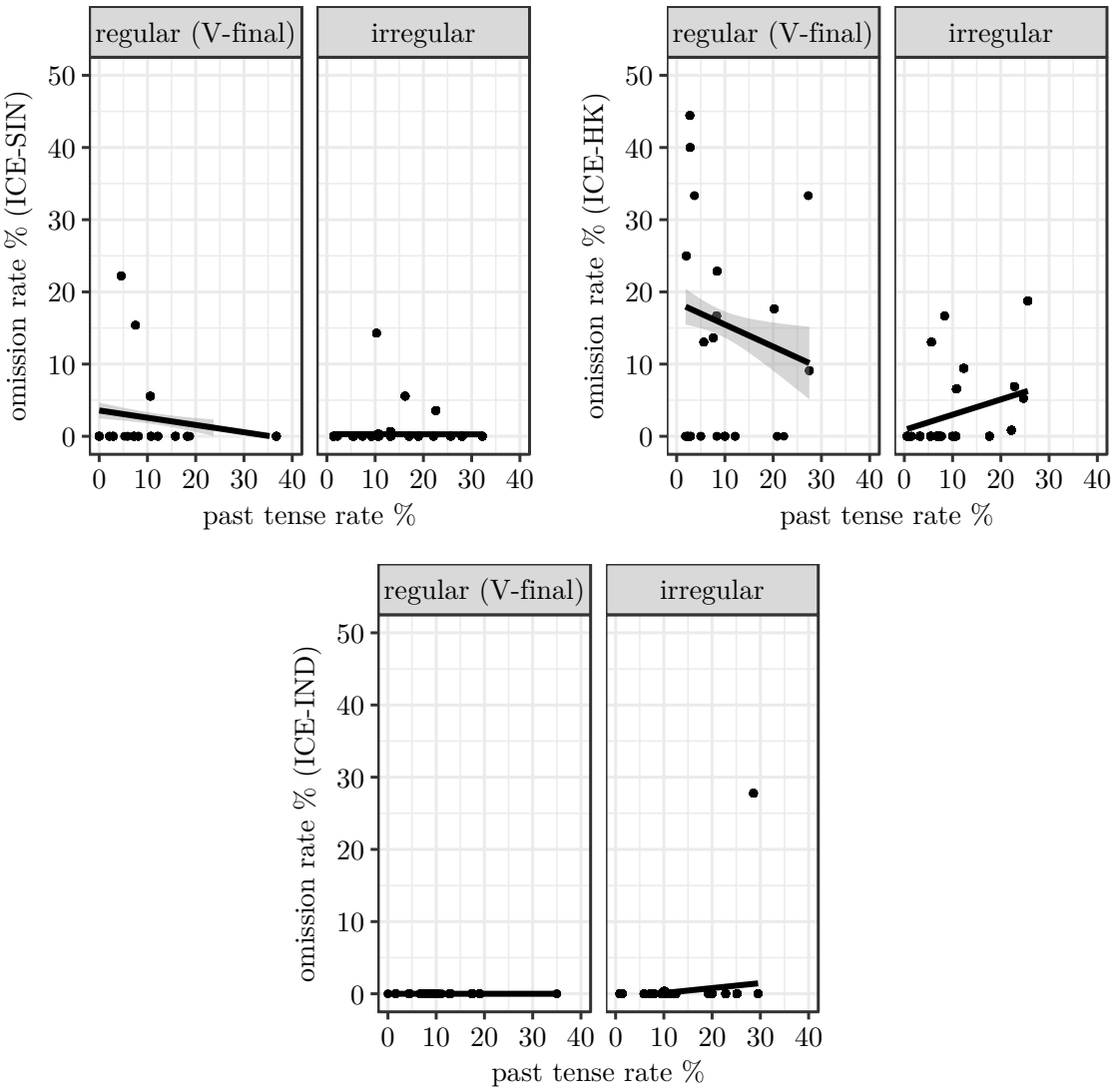

Figure 5.5: Omission rates in ICE by past tense rate, by corpus (functional uses only)

In contrast with the logarithmically transformed relative lemma token frequencies in figure 5.4, the past tense rates were not logarithmically transformed. The relative lemma token frequencies derived from GloWbE are little telling because they are very small, which is why they were logarithmically transformation for visualization purposes (figure 5.4). In contrast, the past tense rates were derived from ICE, and here the actual percentages on the x-axis are telling. Since logarithmic transformation does only change the scale of the x-axis (by distributing the dots more evenly

along the x-axis) but not the trends indicated by the regression lines, it is valid to compare the trends in figure 5.5 to those in figure 5.4.

Compared with figure 5.4, the trends change for all varieties and nearly all verb types; the exception being irregular verbs in ICE-SIN and regular verbs in ICE-IN. A number of clear outliers explain the trends in ICE-SIN and ICE-IND. In ICE-SIN, the respective lemmata are *allow*, *stay*, and *agree*, in ICE-IND it is the irregular verb *forget*.[61] The more interesting case is ICE-HK, and here in particular the regular verbs. When the frequency of occurrence of a regular verb in a past tense context is accounted for exclusively, the expected trend emerges: Infrequent regular verbs are more affected by omission than frequent regular verbs (see hypothesis 1a, sections 1.3 and 5.2). Among the irregular verbs, verbs of higher past tense rate are affected more strongly by omission, but a few outliers define the upward trend.

Since the sampled vowel-final regular verbs and irregular verbs make use of different morphological processes to mark for past tense, it makes sense to account for the impact of type frequencies on the observed omission rates. While figure 5.1 in the previous section provided an overview of the observed omission rates by morphological process, the relative frequency of the respective morphological process was not taken into consideration. Figure 5.6 depicts the observed omission rates by relative type frequency of the morphological processes involved for ICE-SIN and ICE-HK. ICE-IND is not displayed because of the low omission rates observed there.

The relative type frequencies were arrived at as follows: On the basis of the lemma frequency lists extracted from the GloWbE offline database (see section 5.2 for details), lemmata of a token frequency of at least 1,000 (618 verbs in GloWbE SG and 611 verbs in GloWbE HK) were coded for the morphological process underlying their past tense formation. The relative type frequency of each lemma was calculated by dividing the absolute type frequency by the number of verbs considered. The verb *show* was excluded from the analysis because its past participle is either *showed* or *shown*. As pointed out in section 5.2, the samples of vowel- and consonant-final regular verbs and irregular verbs had been drawn from the group of lemmata above the same threshold because lemmata of a frequency of less than 1,000 in GloWbE are unlikely to occur in ICE with sufficient frequencies.

Two aspects are worth pointing out. Firstly, the [t,d] affixation verbs outperform the other verb types in terms of relative type frequency by far across corpora. Sec-

[61] The omission rates (marked for past tense: not marked for past tense) are as follows: for ICE-SIN, *allow* 22.22 (7:2), *stay* 15.38 (11:2), and *agree* 5.56 (17:1); for ICE-IN, *forget* 27.78 (13:5).

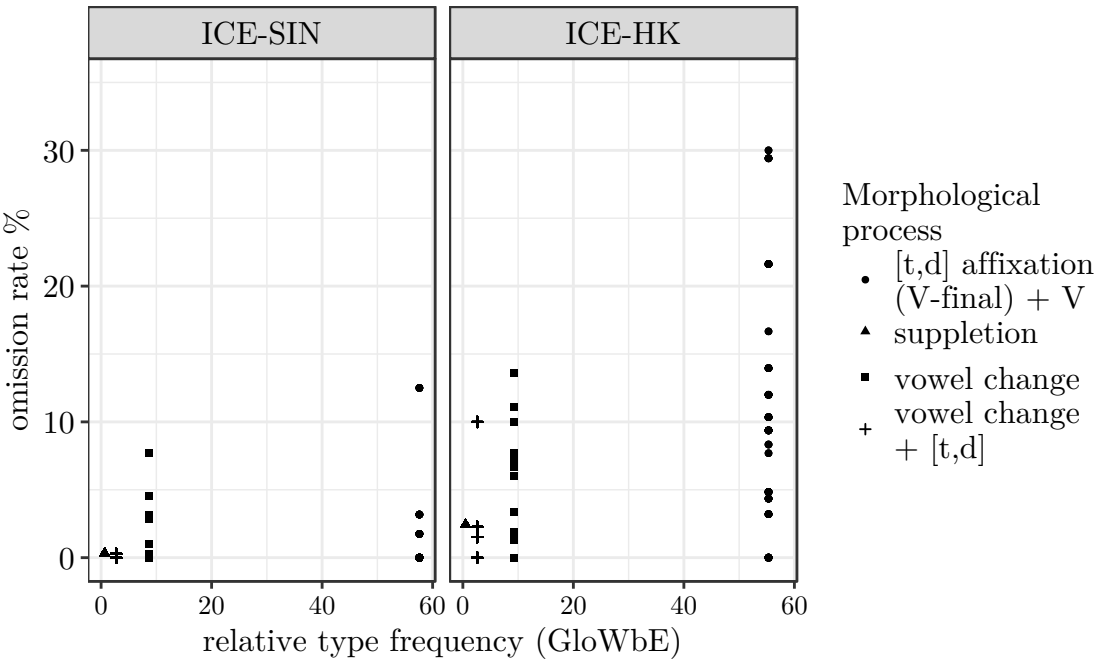

Figure 5.6: Past tense omission rates by relative type frequency (GloWbE) and morphological process, by corpus (formal uses)

ondly, omission rates in [t,d] affixation verbs tend to be higher than those in the other verb types in HKE, but omission in irregular verbs is relatively frequent as well. Omission rates in ICE-SIN are too small to speak of more than chance observations.

5.5 Substratum transfer and institutionalization as constraints on usage frequency

In the previous section we observed that irregular verbs are less prone to omission than regular verbs across corpora. Even relatively infrequent irregular verbs in ICE-HK and ICE-SIN (i.e., irregular verbs of comparable frequency to regular verbs) are less affected by omission than regular verbs. When regular and irregular verbs are considered separately, no clear frequency patterns are observable though. The following paragraphs deal with the question whether substratum transfer and institutionalization account for lack of past tense marking in those cases instead.

Substratum transfer

It was hypothesized that substratum transfer functions as a constraint on frequency effects in case omission and regularization occur considerably more in SgE and HKE

than in IndE irrespective of lemma token frequency. Obviously, the observed patterns of past tense omission need to be explicable by common substrate influence.

Substratum transfer is likely to account for the comparatively high omission rates in environments where the /t,d/ suffix of regular verbs is prone to consonant cluster reduction. This is particularly obvious in ICE-SIN and ICE-HK. While past tense omission in ICE-SIN is nearly exclusively restricted to consonant cluster environments, in ICE-IND only a few outliers point towards a reduction of word-final consonant clusters (cf. section 5.3 and figure 5.1 therein). Those outliers are in stark contrast to the otherwise near absence of omission of verbal past tense marking. Consonant cluster reduction has been reported for all three varieties (e.g., Gut 2009a: 272–276 for SgE; Matthews & Yip 1994: 19 for HKE; Wiltshire 2013 for IndE), but the lack of availability of the original ICE recordings makes in-depth investigations of the impact of consonant cluster reduction on omission rates in ICE impossible. Wiltshire's (2013) finding that consonant clusters in IndE are reduced more by speakers of Tibeto-Burman languages than by speakers of Indo-Aryan languages would be worth investigating on the basis of the speaker background ICE-IND provides.

The fact that the main substrate languages of HKE, SgE, and IndE are aspect- rather than tense-prominent is also worth considering here. As described in section 5.1, Sharma (2009) reports more omission in imperfective than in perfective contexts both among speakers of SgE and among speakers of IndE. The SgE speakers, whose patterns of past tense omission Sharma takes from M. L. Ho (2003), mark 56.2 percent of all verbs in perfective contexts, 36.9 percent of all verbs with lexical stative meaning, and 14.7 percent of the verbs in habitual contexts overtly for past. The IndE speakers (speakers of Hindi) mark 76.6 percent of the verbs in perfective contexts, 44.2 percent with lexical stative meaning, and 29.5 percent in habitual contexts overtly for past. Sharma (2009) explains the predominance of past tense marking in perfective contexts with perfectivity marking in the substrate languages Mandarin and Hindi, which indicate perfectivity by means of the verbal particle *le* in Mandarin and the perfectivity marker *-(y)a* in Hindi, respectively. (Standard) English marks all verbs in a past tense context, irrespective of aspectual distinctions (ibid.: 176). Let us reconsider Sharma's (2009) reattachment scheme:

(5.10) Hindi: [PERF *-a*] → IndE: [PERF *-ed*]

(5.11) Chinese: [PERF *le*] → SgE: [PERF *-ed*]

Speakers of Hindi and speakers of Chinese transfer perfectivity marking to English "whereby the semantic component of a form-meaning pairing in the L1 is re-attached to an L2 form," as Sharma (2009: 179) points out. The process is only depicted for regular verbs though; verbs that form their past tense in an irregular way are not elaborated on. Traditional models of language assume a straightforward structural relationship in regular verbs, whereby the past tense of regular verbs is formed by adding the *-ed* suffix to the bare verb lemma (see section 2.2). Irregular verbs are "listed in the lexicon" (Croft & Cruse 2004: 292) instead. Applied to Sharma's (2009) reattachment scheme, this means that Hindi and Mandarin perfectivity markers promote the attachment of the *-ed* suffix to the base form of regular verbs in perfective contexts, whereas irregular past tense forms are stored as such in the lexicon of learners of English.

Sharma (2009: 175) raises an important issue by pointing out that the SgE speakers analyzed in M. L. Ho & Platt (1993) and the speakers represented in ICE-SIN speak different substrate languages. As we learned in section 3.2, the role of Mandarin in Singapore has strongly increased because of attempts of the post-independence government to promote Mandarin as the shared dialect among ethnically Chinese Singaporeans (compare also Sharma 2009: 175). Since English began to spread in Singapore long before independence, Hokkien, Teochew, Cantonese, as well as the former Hokkien-influenced lingua franca Bazaar Malay have likely influenced SgE in earlier decades (e.g., M. L. Ho & Platt 1993: 9, 27; Ansaldo 2004; L. Lim 2007: 452–453; all in Sharma 2009: 175). As Sharma (2009) points out, "Mandarin may therefore only be an important substrate for the newer ICE data" (175), whereas the speakers in M. L. Ho & Platt (1993) "are less likely to be native speakers of SgE or of Mandarin (Ansaldo 2004; L. Lim 2007)" (Sharma 2009: 175). Like Mandarin, Hokkien, Teochew, Cantonese, and Malay have perfectivity markers (Hokkien *tja*, Teochew *do*, Cantonese *gan*, and Malay *sedang*) but no markers for non-progressive or habitual imperfective contexts with the exception of optional uses of Teochew *do* (non-progressive), Cantonese *zyu* (non-progressive), and Cantonese *hoi* (habitual). Table 5.5 is adopted from tables 1 and 2 in Sharma (2009: 176–177) and summarizes the key aspect distinctions in English, Hindi, and in the substrate languages of SgE.

Note that Hindi indicates non-progressive and habitual aspect by means of *-ta*. According to Sharma (2009) "[t]his means that IndE speakers have a pervasive substrate pressure to mark imperfectivity overtly" (185). The small rates of verbal past tense omission in IndE compared to SgE in both Sharma's (2009) study and

Table 5.5: Key aspect distinctions in English, Hindi, and in the substrate languages of SgE (adopted from Sharma 2009: 176–177)

	English	Hindi	Mand.	Canton.	Teochew	Hokkien	Malay
PAST:							
perfective	—	-(y)a	le	tso	lio	liau	sudah
neutral	-ed	—	—	—	—	—	—
IMPERFECTIVE:							
non-progressive	—	-ta	(-zhe)	(zyu)	(do)	—	—
habitual	—	-ta	—	(hoi)	—	—	—

Note: Progressive aspect is not included here.

the results presented here account for the fact that both perfective and imperfective marking are obligatory in Hindi, but only perfective marking by means of isolated markers is necessary in the SgE substrate languages. This also goes in line with the observation that omission rates in ICE-IND are lower than those in ICE-HK, Cantonese being the main substrate of HKE (see section 3.3.2).

Let us have a look at the extent to which the inflectionally unmarked verbs identified in the corpus analyses presented here occur in perfective and in imperfective contexts. In line with Sharma (2009: 178), clausal perfectivity was identified by means of lexical aspect (telic verbs that describe an endpoint) and by means of additional aspect markers such as perfectivizing adverbials (e.g., *in two minutes*, *all of a sudden*). Table 5.6 depicts the number of inflectionally unmarked verbs in perfective and imperfective contexts whose past tense marking is not prone to phonetic reduction. It focuses on verbs that should be marked for simple past (therefore excluding verbs that are only formally marked for past tense), and both vowel-final regular verbs and irregular verbs are considered.

Table 5.6: Inflectionally unmarked verbs in perfective and imperfective contexts, by corpus (vowel-final regular verbs + following V, irregular verbs, functional uses)

corpus	perfective context	imperfective context	sum
ICE-SIN	2 (28.57 %)	5 (71.43 %)	7 (100 %)
ICE-HK	15 (24.59 %)	46 (75.41 %)	61 (100 %)
ICE-IND	1 (16.67 %)	5 (83.33 %)	6 (100 %)

Lack of verbal past tense marking clearly prevails in imperfective contexts, which was expected based on Sharma's (2009) findings. Regarding the small numbers of unmarked verbs in ICE-SIN and ICE-IND, it is worth recalling that omission in ICE-SIN is largely restricted to phonetic environments that favor consonant cluster reduction, whereas past tense omission hardly occurs in ICE-IND irrespective of phonetic environment and morphological process applied. 5.12 and 5.13 are examples of past tense omission in imperfective (5.12) and perfective (5.13) contexts, respectively.[62]

(5.12) He actually *begin* as a director of dramas such as English costume variety in Sense and Sensibility (ICE-HK:S2B-033#36:1:A)

(5.13) I like was very shocked like he really went to the room and *take* out stuff and like showed us like eh (ICE-SIN:S1A-097#259:1:B)

Let us turn to speaker background information ICE provides us with next. Table 5.7 shows the relevant background information for the speakers who omit functional past tense marking in ICE-HK and ICE-IND. For ICE-SIN, respective speaker background information is not available. From a substratist perspective, birthplace, mother tongue, and overseas experience (which is equivalent to the use of English in international settings) are of particular interest. Age, gender, and the level of education complete the picture.

The homogeneity of the speaker population represented in ICE-HK is remarkable. 53 (out of 61) unmarked verbs were produced by speakers born in Hong Kong and 56 unmarked verbs by speakers who indicated that Cantonese is their mother tongue. Of the speakers who reported China as their birthplace and/or Chinese or Mandarin as their mother tongue, all except for one (for whom no background information is available) received primary and/or secondary education in Hong Kong, which is why they were not excluded from the analyses. Those speakers are not marked as extra-corpus speakers in the ICE-HK transcripts.

The majority of the unmarked verbs in ICE-HK were uttered by females (28 verbs) and by speakers aged between 21 and 25 (36 verbs). The number of unmarked verbs decreases with increasing age, the exception being the age range 41 to 45 with seven unmarked verbs. If omission occurred more systematically in ICE-HK, the predominance of verbal past tense omission in younger speakers would be a sign

[62] Unmarked *take* in 5.13 was not interpreted as an instance of narrative present because all other verbs in the utterance are marked for past tense.

Table 5.7: Background of speakers who omit inflectional marking for simple past

	ICE-HK (61†)	ICE-IND (6†)
age	21–25: 36, 41–45: 7, 17–20: 5, 31–35: 4, 51–55: 3, 46–50: 3, 36–40: 2, 26–30: 1	above 50: 3, 34–41: 1, 26–33: 1, 18–25: 1
gender	female: 38, male: 23	female: 1, male: 5
birthplace	Hong Kong: 53, China: 6, no info: 2	no info: 3, Patna, Bihar: 2, South Kanara, Karnataka: 1
mother tongue	Cantonese: 56, Chinese: 3, Mandarin: 1, no info: 1	Hindi: 2, Kannada: 1, Konkani: 1, Punjabi: 1, Tamil: 1
level of education	University: 32, Secondary: 28, no info: 1	Doctorate: 3, no info: 2, Graduate: 1
overseas experience	10 (no info: 51)	0 (no info: 6)

†Number of verbs lacking verbal past tense marking (vowel-final regular verbs + following V and irregular verbs, functional uses only). Speakers may be represented multiple times per cell depending on the number of unmarked verbs they produced.

of the emergence of a conventionalized feature. However, the assumption has to be interpreted with caution because for reasons of feasibility only the unmarked verbs were coded for speaker information but not the large number of marked ones. Coding the latter would be a necessary step in order to learn about the relative occurrence of unmarked forms in the different age groups. Only ten out of the 61 unmarked verbs were produced by HKE speakers with overseas experience.

The few unmarked verbs in ICE-IND are clear signs of a functionally stable past tense marking system in ICE-IND. Unsurprisingly, the few speakers who omitted past tense marking reported a comparatively large variety of birthplaces and mother tongues. Two of the six unmarked verbs were uttered by a 50+ year old male speaker from Patna, Bihar with Hindi mother tongue background. The other 50+ year old male speaker comes from South Kanara, Karnataka and speaks Kannada, a Dravidian language which marks both tense and perfective as well as imperfective aspect inflectionally. The same is true for Konkani (Indo-Aryan), Punjabi (Indo-Aryan), and Tamil (Dravidian). The levels of education both the ICE-HK and the ICE-IND speakers reported show that the corpus data were collected among speakers with high educational attainment.

We can conclude that influence from isolating substrata certainly accounts for the differences in omission rates between HKE and IndE. The main substrata of both HKE (Cantonese) and IndE (Hindi) are aspect-prominent, but omission rates in IndE are too low to be explicable by more than accidental omission. In ICE-SIN, omission is nearly exclusively restricted to phonetic environments that favor consonant cluster reduction, and the same is true for ICE-HK, although the trend is less clear in HKE. Verbs along the frequency cline are affected, which is why substratum transfer functions as a constraint on potential frequency effects.

Institutionalization

Finally, let us elaborate on the question whether institutionalization functions as a constraint on frequency effects. This was hypothesized to be the case should patterns of past tense omission be particularly stable in SgE, the most institutionalized varieties considered, irrespective of lemma token frequency (see hypothesis 3, sections 1.3 and 5.2).

The degree of institutionalization of the varieties of interest was elaborated on in chapter 3 and determined on the basis of Schneider's Dynamic Model (cf. Schneider 2007: 35–36, 67). SgE is in stage 4 ("endonormative stabilization") and, according to Ooi (2001a), even on its way to stage 5 ("differentation") because variety-specific features have become increasingly conventionalized; visible in particular in the development of the local vernacular Singlish. HKE and IndE are in stage 3 ("nativization"), but Mukherjee (2007) argues that IndE is progressing to stage 4 because ongoing protests against Hindi in the south have resulted in the "reconstruct[ion of] a radically new, locally based identity" (Schneider 2007: 49).

In the development of hypothesis 3, Van Rooy's (2011) distinction between learner errors and conventionalized innovations was elaborated on (see also section 2.4). Based on this distinction, it can be assumed that past tense omission only constitutes a conventionalized innovation in case the feature is (grammatically) stable. When omission does not occur systematically, it is likely a learner feature. SgE as the most institutionalized variety considered should show the highest degree of stability. If institutionalization constrains frequency effects, this should be the case irrespective of lemma token frequency.

In ICE-SIN, omission of verbal past tense marking in the 60 sampled verbs is nearly exclusively restricted to environments in which the /t,d/ suffix is likely affected by consonant cluster reduction. Neither lemma token nor type frequency can explain the observed omission rates (see section 5.4). Thus, while the clear restric-

tion of omission to consonant cluster environments is a sign of the high degree of institutionalization of SgE, institutionalization does not function as a constraint on frequency due to the fact that frequency does not play an important role; at least when we focus on environments that are not prone to consonant cluster reduction.

For HKE, the picture looks different. While omission is comparatively strong in consonant cluster environments, the phenomenon occurs in other phonetic environments and in irregular verbs as well. Omission takes place across the frequency range (recall the expected downward trend when past tense rates are considered though) and is neither restricted to certain speakers nor particularly strong in the presence or absence of a time adverbial. All these factors are clear indicators that omission of verbal past tense marking is very much a learner feature of HKE that does not follow clear patterns. There is one exception though: Irregular verbs (that learners of English presumably pay particular attention to) are less prone to omission than regular verbs. However, HKE irregular verbs are more affected by omission than irregular verbs in IndE and SgE. The fact that the majority of the unmarked verbs were produced by young speakers (see above) is an observation worth mentioning but cannot be interpreted as a sign of conventionalization of the feature in younger generations given the lack of systematicity with which omission occurs.

In ICE-IND, omission was hardly found in the 60 sampled verbs with the exception of a few outliers that were detected across the different verb types examined. Those outliers clearly constitute learner errors just as those in HKE, but they are not promoted by isolating substrates (as in the case of HKE). The substratum effect that Sharma (2009: 177–179) observes in her data and which she explains with the aspect-prominence of Hindi (resulting in higher omission rates in imperfective than in perfective contexts), could not be replicated. Crucially, "most [of Sharma's speakers] are small shop owners, shop employees, or are unemployed" (ibid.: 174), whereas the speakers recorded for ICE-IND have reached high levels of education. Education and social status might play a role here, and a contrastive analysis of speakers that differ in this regard would be a follow-up study worth conducting.

5.6 Concluding remarks

To conclude, the previous sections have shown that omission rates in the contact varieties of interest can be best explained by means of a combination of frequency effects, substratum transfer, and institutionalization. Across varieties, omission rates

turned out to be surprisingly small, but a clear difference was observed between the omission of the /t,d/ suffix in regular verbs and lack of past tense marking in irregular verbs. Even a focus on regular and irregular verbs of comparable frequency revealed that regular verbs which are not prone to consonant cluster reduction are more affected by omission of verbal past tense marking than irregular verbs. The high frequency of use and salience of verbs with irregular past tense marking is likely to account for the reported differences. Both substratum transfer and institutionalization additionally help to explain the observed omission rates.

The low omission rates across varieties might be explicable by the types of speakers recorded for ICE. All spoken sections in ICE were taken into consideration, which resulted in a more heterogeneous speaker group than if only, say, face-to-face conversations had been accounted for. While the speech contents across sections in ICE-HK and ICE-IND are manifold (the metadata for ICE-HK provide much more detailed information than those for ICE-IND), most speakers have reached a high level of education. The majority of the ICE-HK speakers have attended university. Levels of education the ICE-IND speakers provided start with college, but about half of the speakers did not give information about their educational attainment. Unfortunately, the ICE-SIN metadata are not available, so the background of the speakers recorded for the spoken part of ICE-SIN could not be elaborated on. Skimming through the corpus files of all ICE corpora considered left the impression that even the face-to-face conversations adopt a certain formality level; particularly those collected in ICE-IND.

6 Omission of inflectional noun plural marking: A corpus-based account

This chapter presents a case study on omission of nominal plural marking in the three Asian varieties of interest. The design of the case study parallels that of the case study on omission of verbal past tense marking presented in chapter 5; with the exception that nouns with regular plural marking are focused on exclusively. The reason is that nouns with irregular plural marking constitute a relatively small class with too few forms of sufficient frequency to work with (see section 6.2 for details). As pointed out in section 1.2, both nouns that are accompanied by a determiner with plural reference (e.g., a quantifier or a numeral) and nouns that are not preceded or followed by a respective determiner are accounted for. Compare examples 6.1 (no determiner) and 6.2 (determiner), respectively. The nouns *question* and *book*, which lack their plural suffix *-s*, are in italics.

(6.1) Uhm <„> so uh if you look at uhm my my writing uhm for the representation the dots denote other possible slots that answer *question* like how why instrument associated et cetera as suggested by the case grammar (ICE-HK:S2B-048#69:1:A)

(6.2) He translated many *book* <,> of that Indo-Persian books <,> and he used to read them <„> (ICE-IND:S1B-005#101:1:S)

A usage-based account of nominal plural marking is of particular interest here, and the role of substratum transfer and institutionalization are additionally considered. The analyses are conducted on the basis of the spoken part of ICE. An additional investigation of omission rates in GloWbE is supposed to reveal whether the phenomenon has found its way into language use on the web.

Section 6.1 presents previous research on omission of nominal plural marking, followed by section 6.2, which introduces the sample of nouns worked with and elaborates on the hypotheses. Section 6.3 continues with a general overview of the observed omission rates, whereas section 6.4 offers a usage-based account of omission

of nominal plural marking. Section 6.5 focuses on the impact of substratum transfer and institutionalization on omission rates and investigates whether those factors constrain potential frequency effects. Section 6.6 concludes.

6.1 State of the art

Omission of inflectional noun plural marking is particularly well researched for SgE compared with the other two varieties of interest. According to eWAVE, plural marking is generally optional for nouns with human and non-human referents in CSE and HKE (eWAVE; features 57 and 58). In IndE, in contrast, omission of inflectional noun plural marking exists but is extremely rare both with human and non-human referents. However, it should be kept in mind that eWAVE lists typical or salient features rather than necessarily frequent ones. The following paragraphs provide a summary of accounts of plural omission in the three varieties of interest.

Singapore English

Lack of plural marking is a feature often described as typical of SgE, but few sources have investigated the phenomenon empirically. An in-depth account of the absence of plural marking in SgE is provided by Ziegeler (2015), who examines the feature in CSE[63] from the perspective of Construction Grammar. Ziegeler treats lack of plural marking as one formal characteristic of the so-called "bare noun phrase construction," bare nouns being defined as nouns that are not marked for number or that are not preceded by determiners (ibid.: 182). On the formal side, the bare noun phrase construction is characterized either by "the absence of the determiner on singular count nouns or the absence of plural marking (zero-plural) on plural count nouns" (ibid.: 181–182), i.e., it "has the form of a mass noun" (ibid.: 182). On the functional side, bare noun phrases are non-specific, i.e., they do not refer to specific entities. Ziegeler stresses that in-depth research on plural marking in general and in the context of article omission in particular is lacking (ibid.).

Ziegeler (2015) is particularly interested in "the semantic characteristics which govern [the] absence [of the plural suffix] in particular environments, such as in reference to an indeterminate quantity" (183). This is why her study places particular weight on the role determiners that accompany nouns with plural reference play for plural omission. In the advertising literature she investigates, Ziegeler observes that plural marking is likely to occur with premodifying quantifiers in Singlish, but she

[63] Ziegeler (2015) refers to CSE as "Singapore Colloquial English."

describes the phenomenon as not "entirely rule-governed" (ibid.: 201–202). In fact, the role quantifiers or other determiners preceding the noun play for plural marking, or lack thereof, has been examined in a number of accounts of (lack of) plural marking in SgE with contradictory findings.

Alsagoff & C. L. Ho (1998: 144) note that plural affixes are relatively unlikely to be omitted in the presence of premodifying quantifiers, such as *many*, *few*, or preceding numerals. This view is challenged by Wee & Ansaldo (2004: 64), who provide evidence of plural omission after quantifiers or numerals from the *Grammar of Spoken Singapore English Corpus* (GSSEC; cf. L. Lim 2004; L. Lim & Foley 2004), such as in *ten thousand of my friend*. Deterding (2007) mentions an example of lack of plural marking after the quantifier *few* as well, namely in the speech of a Chinese-origin female university undergraduate of ethnically Chinese origin (*I mean the few country that I've been to are . . .*). However, he states that "the overwhelming majority of nouns [in her speech] have the standard plural marking" (44). M. L. Ho & Platt (1993: 22) report findings by M. L. Ho (1981) on the speech of 50 ethnically Chinese English-medium-educated speakers of SgE of different levels of education. They observe that countable as well as uncountable nouns are marked for plural in particular when a determiner precedes them and less so if no determiner is present. In an even earlier account of SgE as spoken among Nanyang University[64] graduates, Elliott (1983: 52–53) notes the following:

> It is not possible to say whether the endings are omitted more often when they are redundant. The impression gained in listening to [the graduates'] speech is that they ignore any thought of whether the object they are discussing is one or many.

This is a remarkable observation that deserves further attention and that is directly linked to the lack of the conceptualization of plural versus singular entities in the main substratum languages of SgE (see section 6.5 for details).

With reference to SgE and (other) L2 varieties of English, Biewer (2015) points out that "cognitive principles of speech production, SLA and substrate influence play a role in the emergence of [inflectionally unmarked nouns with plural reference]" (172). Cognitively, "perceptually more conspicuous" plurality markers, i.e., quantifiers and numerals are likely to be chosen over less salient ones, i.e., the plural

[64]Nanyang University and the University of Singapore merged to the National University of Singapore in 1980. Nanyang University was largely Chinese-medium, the University of Singapore English-medium.

suffix *-s*. Additionally, substrate influence is likely to push the use of analytic plurality markers, although the plural suffix is more frequent in English than preceding quantifiers and numerals. Particularly in the early stages of the learning process, learners opt for salient markers at the expense of seemingly redundant ones (cf. Sand 2005: 181; Williams 1987: 176). Biewer (2015) provides examples from Cook Island English, Samoan English, and Fiji English, where "plurality is marked by a plural numeral or quantifier but not by plural *-s* on the noun" (173). Substratum transfer occurs insofar as free morphemes in premodifying position can be used to mark plural in Oceanic languages such as Cook Islands Maori, Samoan, and Fijian (see ibid.: 173 for details). While HKE and IndE are not explicitly mentioned, Biewer's (2015) arguments are worth keeping in mind for these varieties.

The phonetic environment is less an issue for omission of nominal plural marking than it is for omission of verbal past tense marking. L. Lim (2004: 33) mentions that voiceless alveolar fricatives in SgE are sometimes deleted when they occur in the final position in a consonant cluster and when they are preceded by /n/, /t/, or /k/ (as in *license* or *relax*). At the same time, reduction of the penultimate consonant (instead of the final one) is possible, as in *that's*, *facts*, *parents*, *depends*. Gut (2005) investigates two-consonant clusters of the types "plosive+/s,z/ (/ts, ks, ps, dz/), nasal+plosive (/nt, nd, ŋk, md, mp/), /l/+plosive (/ld, lt/), plosive+plosive (/pt, kt/), /f, v/+plosive (/ft, vd/) and /s,z/+plosive (/st, sk, sp, zd/)" (20–21) and finds the least reduction in plosive+/s,z/ clusters. In 43.5 percent of all cases both the plosive and /s/ or /z/ are retained, and in 44.7 percent of all cases the /s/ or /z/ is retained. Gut (2005) additionally looks at 43 three-consonant clusters of the type lateral/nasal+plosive+/s/ (/ndz, nts, ldz/) and seven three-consonant clusters with "two plosives and an /s/ in various positions" (22). In four percent of the cases all consonants are retained, in 59 percent one consonant is deleted, and in 37 percent two consonants are deleted. When one consonant is deleted, it is always the plosive. In 89 percent of all cases, the /s/ is retained and in eleven percent the nasal. These findings show that often sounds other than alveolar fricatives are omitted in final consonant clusters.

Hong Kong English

There is relatively little research on omission of inflectional noun plural marking in HKE. One notable exception is the work by Setter et al. (2010), who describe uses of the plural suffix as seemingly random at first glance. Similar to Ziegeler's (2015) account of the bare noun phrase construction in SgE, the authors stress the

equivalence in form between plural countable nouns that lack plural marking and singular countable nouns that are used without prenominal modifiers such as articles (cf. Setter et al. 2010: 45). Whether the plural reference is intended by the speaker often has to be inferred from the context or remains unclear. In case a determiner with clear plural reference is present, it can be accompanied by a noun that lacks inflectional plural marking (ibid.: 46). Budge (1989), in contrast, proposes that "HKE speakers tend to mark plural where there is some semantic reminder that the noun is to be marked as plural" (41). She distinguishes between markers that are "neutral with respect to plurality" (ibid.: 39) (e.g., *other* or *certain)*, clear plurality markers (e.g., *one of the*), and the modifiers *any* and *some*. She notes that plural marking is particularly likely "the stronger or more unambiguous" (ibid.: 41) such a reminder is, i.e., with clear plurality markers. For Budge (1989), "this runs contrary to any expectation that speakers will omit plural marking after plural indicating modifiers because of redundancy" (41). Recall Biewer's (2015) argument on substratum influence above, according to which influence from substrate languages promotes the use of analytic plurality markers at the expense of seemingly redundant synthetic markers in the early stages of the learning process.

Budge (1989) also briefly elaborates on the impact of the phonetic environment on omission of nominal plural marking. Cantonese has no syllable-final consonant clusters (compare also Matthews & Yip 1994), which can lead to pronunciation difficulties that affect the plural suffix (cf. Budge 1989: 42). The same argument has been made for speakers of SgE, but research by L. Lim (2004: 33) and Gut (2005: 20–21) shows that, in contrast with plosives, alveolar fricatives in word-final consonant clusters are little affected by reduction. Unfortunately, Budge (1989: 42) provides no information on the difference in omission rates between nouns in which the plural suffix is part of a consonant cluster and nouns for which this is not the case. Many of the examples she provides are nouns that end in an alveolar plosive. Respective nouns are a special case because by adding the plural suffix they add an extra syllable.

As regards substratum influence from Cantonese, Budge (1989: 43–44) notes the following:

> If Cantonese has influenced HKE, it has done so in a general way. Just as the pre-nominal elements, especially numerals in Cantonese, [sic] serve to indicate to the speaker that the following noun has plural reference, so the pre-nominal elements in English indicating plural serve to signal

> to the HKE speaker that the noun should be marked for plural. Thus only indirect influence from Cantonese can be discerned.

Consequently, Budge (1989) reasons that plurality markers remind HKE speakers to mark their nouns for plural. As pointed out above, she sees evidence against the redundancy of inflectional marking after plurality markers in her data insofar as inflectional marking for plural is particularly likely after markers that unambiguously indicate the plural reference. Unfortunately, Budge (1989) does not elaborate on the reason(s) for considering the familiarity with pre-nominal elements from Cantonese an indirect influence on plural marking only. Section 6.5 on the role of substratum transfer for omission of nominal plural marking deals with the matter in more detail.

Indian English

Accounts of plural marking in IndE mainly focus on uses of uncountable nouns as countable nouns (see section 7.2) and vice versa as well as on article omission, whereas detailed investigations of potential omission of the plural suffix are lacking. Sharma (2005) elaborates on article use in the English of twelve first-generation adult Indian immigrants in California, whose language data were gained by means of interviews. She observes that both transfer from the speakers' L1 and discourse pragmatic principles account for patterns of article use among her subjects (cf. Sharma 2005: 563). As the previous sections on SgE and HKE showed, the equivalence in form of plural countable nouns that lack plural marking and of singular countable nouns that are used without prenominal modifiers like articles often makes it necessary to consult the context to understand the reference. If this is an issue in IndE as well, the matter has not been addressed, to the author's knowledge.

6.2 Sample choice and hypotheses

Sample choice

As the previous section showed, a number of factors have been identified in the literature that coincide with or impact omission of nominal plural marking, among them non-standard bare noun phrases which can either result from a missing plural suffix or from a missing article. In contrast with omission of verbal past tense marking, where consonant cluster reduction clearly impacts omission of the past tense suffix (compare section 5.3), the plural suffix is less affected by consonant cluster reduction. Plosives that precede the plural suffix constitute a potentially tricky case,

which is why nouns ending in a plosive where the plosive is not part of a noun-final consonant cluster are not considered here (for details see section 6.1).

Samples of nouns with regular plural marking were investigated in ICE and GloWbE rather than focusing on subsets of the corpora and identifying all unmarked nouns therein for the reasons mentioned in section 5.2 for past tense omission. This approach guaranteed higher token numbers to work with, reduced the risk of speaker bias, and allowed for investigation of register differences in ICE. A comprehensive account of plural omission across lemmata in GloWbE and subsetting the corpus (beyond the broad distinction between blogs and general web-based contents) would not have been possible. All marked and unmarked nouns were retrieved and coded manually.

As in the corpus study on omission of verbal past tense marking, lists of the lemma token frequencies of all nouns in GloWbE were extracted from the GloWbE full-text offline database for each of the varieties separately because GloWbE provides more encompassing frequency rankings than ICE. A comparison of the GloWbE frequencies with the ICE frequencies revealed that nouns of a frequency of less than 8,000[65] in GloWbE are unlikely to occur in ICE at all. The 200 nouns that remained after nouns of a frequency of less than 8,000 in GloWbE had been discarded from the frequency list served as a pool from which a random sample of 15 nouns of varying token frequencies was chosen. Additionally, five nouns (*girl*, *school*, *boy*, *teacher*, *place*) that had been investigated for a preliminary study were added to the sample. Taken together, the sampled nouns are *boy*, *year*, *thing*, *problem*, *friend*, *day*, *hour*, *student*, *eye*, *parent*, *term*, *detail*, *shoe*, *school*, *point*, *way*, *question*, *teacher*, *reason*, and *girl*. A forced binary distinction between frequent and infrequent lemmata was opted against because the frequency cline made it difficult to draw the line somewhere meaningful. The 200 remaining nouns had initially been divided into quartiles, but after discarding the quartile with the lowest token frequencies too few differences in frequency remained to meaningfully distinguish between frequent and infrequent nouns.

The initial idea had been to compare omission rates of nouns with regular plural marking to those of nouns with irregular plural marking. However, a look at the 200 most frequent nouns derived from the GloWbE frequency lists revealed that only three nouns form their plural irregularly (namely *man*, *woman*, and *child*). Even

[65] The respective number of verbs was 1,000 (see section 5.2), which shows the enormous variation in noun compared to verb usage.

going down further in the GloWbE frequency ranking did not provide further nouns with irregular plural marking that occur with sufficient frequency in ICE. Nouns that are not overtly marked for plural (e.g., *sheep*) were not considered for obvious reasons. Consequently, a comparison of omission rates in regular and irregular nouns made no sense here.

As section 6.1 showed, an important point of interest in previous investigations of plural omission in the three varieties of interest has been the role a preceding determiner (or lack thereof) plays for omission of the plural suffix. For the corpus study presented here, a distinction was made between articles, demonstratives, preceding quantifiers, and preceding numerals. A particularly interesting case is the determiner *one of* as in *one of the/those/her books*, which might trigger lack of plural marking because of the numeral *one*. This will be discussed in section 6.3.

Potential hits with corpus-internal markup (e.g., extra-corpus material marked with <X></X> or uncertain transcriptions marked with <?></?>) were discarded from the analyses, as were wrongly marked forms and instances of self-correction. Additionally, it was made sure that plural omission did not occur speaker-specifically. Whenever lack of subject-verb agreement or missing articles made it difficult to determine whether a singular or a plural entity is referred to and the context did not help, respective hits were ignored. The issue of missing articles and their connection to missing nominal plural marking will be taken up again in section 6.3.

Hypotheses

As in the corpus study on lack of verbal past tense marking, the impact of frequencies of use, substratum transfer, and institutionalization on nominal plural marking is of particular interest here.

Regarding usage frequencies, it was hypothesized in line with the Conserving Effect (e.g., Bybee 2007: 10) that frequent nouns are less prone to omission than infrequent nouns because of their relatively strong entrenchment in the speakers' mind. This is backed up by language acquisition research, according to which frequent forms are acquired first and infrequent forms later. The Reduction Effect (e.g., Bybee 2007: 11) does not apply because only nouns that are not prone to consonant cluster reduction affecting the plural suffix are taken into account.

In contrast with the corpus study on omission of verbal past tense marking, only nouns with regular plural marking were considered, which is why only the frequency cline within that group of nouns was of interest. This choice determined the frequency measure used. While lemma token frequencies were accounted for, lemma

type frequency was not of interest because all sampled nouns apply the same type of plural marking. The respective lemma token frequencies were drawn for each of the varieties of interest from the GloWbE frequency lists extracted from the GloWbE full-text offline database. As in the corpus study on past tense omission, frequencies of use were not approximated by means of ICE to guarantee an independent frequency database. Relative frequencies (frequency by corpus size) rather than absolute frequencies were used to account for the differences in corpus size of GloWbE SG, GloWbE HK, and GloWbE IN. Depending on whether omission is a learner feature or a variety-specific innovation (Van Rooy 2011: 192; see section 2.4), plural omission was expected to affect infrequent nouns more or first and frequent nouns less or later (see hypothesis 1a, section 1.3).

Secondly, the choice of varieties enabled accounting for the impact of substratum transfer on omission rates. Of particular interest here is Biewer's (2015) point that "perceptually more conspicuous" (172) markers for plurality, such as quantifiers or numerals, are likely to be chosen over less salient markers, such as the plural suffix *-s*, in the early stages of the learning process in particular. In the previous section we learned that lack of nominal plural marking has been described for both SgE and HKE, and its equivalence in form with singular countable nouns that are used without articles has been stressed in the literature. In IndE, omission of nominal plural marking is not an issue instead. As the following paragraphs will show, speakers of SgE and HKE are not used to inflectional affixes from the substrate languages they speak, meaning they are particularly likely to opt for salient plurality markers. The lacking conceptualization of plural versus singular entities in the main substratum languages of SgE and HKE likely supports this trend as well. Consequently, the following was assumed: In case SgE and HKE have similar degrees of plural omission that can be explained by substratum transfer and differ from IndE, transfer accounts for omission. Substratum transfer is considered a constraint on frequency in case patterns of omission can be explained by transfer but not by frequency effects (see hypothesis 2, section 1.3).

To account for the impact of institutionalization on omission rates, Van Rooy's (2011) account of learner errors versus conventionalized innovations (section 2.4) in New Varieties is considered again. Due to the fact that social factors like acceptance are necessary for the development of conventionalized innovations, plural omission assumably only constitutes a conventionalized innovation in case the phenomenon is (grammatically) stable and accepted. Stability can be assumed in case plural omis-

sion either occurs very frequently or follows clear patterns in the corpus data. The degree of acceptability of nominal plural marking was tested in the perception experiment elaborated on in chapter 8. Patterns of plural omission were expected to be particularly stable in SgE because of the variety's high degree of institutionalization (see hypothesis 3, section 1.3).

6.3 Omission rates: An overview

Omission rates by corpus

Let us have a look at the omission rates in ICE first. As discussed above, only nouns with regular plural marking in which the plural suffix is not prone to consonant cluster reduction were considered. The omission rate of each lemma was calculated by dividing the number of unmarked forms of the lemma with plural reference by the sum of marked and unmarked forms of the lemma with plural reference. Table C.1 in appendix C provides for each sampled noun the number of marked and unmarked forms, the resulting omission rate, and its lemma token frequency in ICE.

Table 6.1 shows the omission rates by corpus including the number of marked and unmarked forms each. Figure 6.1 visualizes the information.

Table 6.1: Rates of plural omission and median values by corpus

corpus	omission rate %	(marked:not marked)	median
ICE-SIN	1.51	(3518:54)	0.93
ICE-HK	7.58	(3172:260)	6.36
ICE-IND	2.28	(3388:79)	1.76
ICE-GB	0.00	(2343:0)	0.00

The omission rates turned out to be surprisingly low across corpora and strikingly low in ICE-SIN compared with ICE-HK and ICE-IND. The largest spread of the middle 50 percent of the data is observable in ICE-HK, while outliers that are more than "1.5 times the interquartile range" (Baayen 2012: 30) are observable in ICE-SIN and ICE-IND only. In ICE-SIN, the clear outlier is *shoe* (omission rate: 7.69, marked: 24, not marked: 2), an infrequent lemma for which few unmarked forms weigh heavily. The outlier in ICE-IND is *reason* (omission rate: 9.43, marked: 48, not marked: 5). Neither of the outliers is speaker-specific, but all five instances of unmarked *reason* are preceded by the numeral *one of* (e.g., *That is the one of the main reason* <„>, ICE-IND:S1A-025#187:1:A).

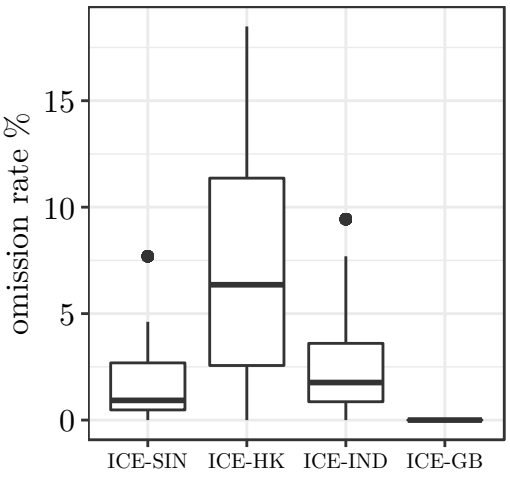

Figure 6.1: Plural omission rates by corpus

Let us turn to GloWbE next. For each of the 20 sampled nouns (see section 6.2), random samples of bare noun forms that potentially constitute nouns without inflectional plural marking were drawn for the three Asian contact varieties. The bare noun forms were obtained by adding the part-of-speech tag ".[n*]" to the noun stem. As in the corpus study of lack of verbal past tense marking in GloWbE (see section 5.3), the sampling function of GloWbE's web-based interface was used to draw samples of 200 in GloWbE SG and GloWbE HK and samples of 400 in GloWbE IN. The omission rates were obtained in the same way as the past tense omission rates. The reader is referred to section 5.3 for details. Suffice it to say that the number of nouns that are inflectionally marked for plural was obtain by means of adding the part-of-speech tag ".[n*]" to the plural form of the noun lemma (e.g., "boys.[n*]"). Figure 6.2 shows the observed patterns, and table C.2 in appendix C depicts the respective numbers for all lemmata of interest.

The plural omission rates are even smaller than the past tense omission rates in GloWbE (compare figure 5.3). While omission of verbal past tense marking is hardly observable either, rates of past tense omission for single lemmata reach 8.81 percent (*finish* in GloWbE HK). Compared with that, omission of nominal plural marking is basically non-existent in GloWbE, which is why it can be deduced that the phenomenon has not found its way into language use on the web. Across lemmata, the absolute number of bare nouns that should be marked for plural in the samples does not exceed three (*shoe* in GloWbE HK). It needs to be pointed out that because of

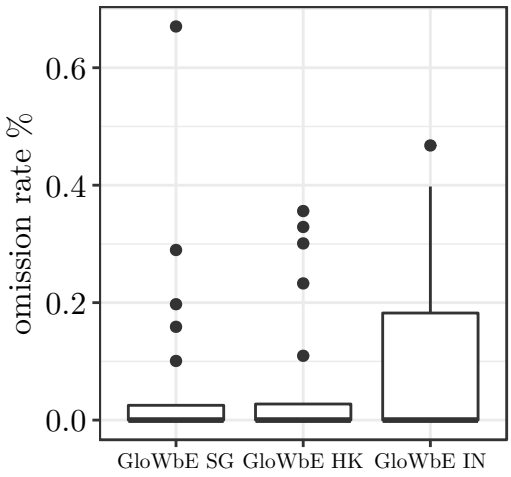

Figure 6.2: Omission rates in GloWbE

the small sample sizes of 200 and 400 single inflectionally unmarked nouns weigh heavily.

Omission rates by syllable number

To gain insights into the impact of word length on omission, syllable numbers were accounted for. The number of phonetic syllables was adopted from WebCelex (Max Planck Institute for Psycholinguistics 2001) for each lemma, but it was not possible to check whether the WebCelex syllable counts match the syllable numbers in the original recordings because the ICE recordings are not available. This is why syllable number had not been accounted for in the sampling process in the first place. Since none of the sampled nouns end in a sibilant, syllable numbers of the singular and plural forms are identical. Table 6.2 summarizes the observed omission rates by corpus and syllable number.

Table 6.2: Plural omission rates by corpus and syllable number

corpus	omission rate % (marked:not marked)			
	1 syllable		2 syllables	
ICE-SIN	1.31	(2,477:33)	1.98	(1,041:21)
ICE-HK	5.37	(2,146:122)	11.86	(1,026:138)
ICE-IND	1.52	(2,330:36)	3.91	(1,058:42)

Omission rates are comparatively high for bisyllabic nouns in ICE-HK. High rates of past tense omission were observable in monosyllabic verbs in ICE-HK instead

(compare section 5.3). Unfortunately, the lack of availability of the original recordings makes it impossible to investigate in more detail whether plural omission is favored with bisyllabic nouns.

Omission rates by determiner type

As pointed out in section 6.1, the role preceding determiners play for plural omission has been of interest in the literature on SgE and HKE in particular, but the findings are contradictory. While some studies find lower omission rates in the presence of determiners such as quantifiers or numerals (e.g., Ziegeler 2015 for SgE; Budge 1989 for HKE), others provide examples of plural omission with preceding determiners (e.g., L. Lim 2004: 64 for SgE; Setter et al. 2010: 46 for HKE).[66]

Table 6.3, adopted from the *Longman Grammar of Spoken and Written English* (Biber et al. 2007: 259), provides an overview of commonly distinguished co-occurrence patterns of different determiner types and nouns. In the *Longman Grammar*, the term "determiner" does not only refer to articles but also to possessives, demonstratives, quantifiers, and numerals. In the literature, the terms "determiner" and "article" are often used interchangeably, but henceforth articles are referred to as "articles," the other determiners as "possessives," "demonstratives," "quantifiers," and "numerals," and all determiner types together as "determiner types" to avoid confusion. Of course, the list of examples provided for each determiner type in the table is far from all-encompassing.

The table makes a distinction between countable and uncountable nouns. While only countable nouns are of interest at this point (see section 7.2 for an account of uncountable nouns that are marked for plural like countable nouns), this distinction is crucial as far as article use is concerned. In English, the choice to use an indefinite or a definite article depends on the reference made, and only in the case of indefinite countable nouns the article unambiguously indicates the singular reference. I.e., the indefinite article *a(n)* necessarily needs to be accompanied by a singular noun. In all other cases reference to the singular or plural is unclear, and the context needs to be considered in the absence of other determiner types. Consequently, when contextual cues are lacking, reference to the plural is only unambiguously made when the noun

[66] Rüdiger (2017) investigates lack of plural marking in English spoken in Korea by means of a self-compiled corpus of spoken language called *SPOKE*. Among other things, she investigates "minus-plural marking" (ibid.: 104), as she calls the feature, after quantifiers and numerals and obtains a "plural redundancy reduction rate" of 29 percent after quantifiers (ibid.: 115) and of 15 percent after numerals (31 percent in the latter case, when time reference nouns like *month* or *year* are accounted for; ibid.: 116–117).

Table 6.3: Co-occurrence patterns of determiner types and nouns commonly distinguished (adopted from the *Longman Grammar of Spoken and Written English*; Biber et al. 2007: 259)

determiner type	countable nouns		uncountable nouns
	singular	plural	
article	—	*books*	*money*
	a book	—	—
	the book	*the books*	*the money*
possessive	*my/your . . . book*	*my/your . . . books*	*my/your . . . money*
demonstrative	*this book*	*these books*	*this milk*
	that book	*those books*	*that milk*
quantifier	*every book*	—	—
	each book	—	—
	—	*all (of) the books*	*all (of) the milk*
	—	*many (of the) books*	*much (of the) milk*
	—	*a great many books*	*a great deal of milk*
	—	*a lot of books*	*a lot of milk*
	—	*lots of books*	*lots of milk*
	—	*plenty of books*	*plenty of milk*
	—	*some (of the) books*	*some (of the) milk*
	—	*(a) few books*	*(a) little milk*
	—	*several books*	—
	—	*a couple of books*	—
	—	*enough books*	*enough milk*
	either book	*both books*	—
	neither book	—	—
	any book	*any (of the) books*	*any (of the) milk*
	no book	*no books*	*no milk*
	—	*none of the books*	*none of the milk*
numeral	*one book*	*two/three . . . (of the) books*	—

is preceded by a quantifier, a numeral, or a demonstrative. The lack of contextual cues and determiner types other than articles made it necessary to exclude potential candidates for which the plural reference could not be determined. Possessives never indicate an existing plural reference of the nouns following them.

A particularly interesting case is the premodifier *one of* as in *one of the books*. Budge (1989) provides an example of plural omission following *one of* for HKE (*an one of my bes(t) frien(d) was study in U, United States* ...; 41) and describes this phenomenon as belonging to a category of "modifiers that contain words which, on their own, accompany only singular nouns, but in which the modifier phrase can only accompany nouns marked for plural" (40). Table 6.4 depicts the observed omission rates in the presence and absence of demonstratives, quantifiers or numerals, i.e., determiner types that clearly indicate an existing plural reference.

Table 6.4: Omission rates in the presence and absence of demonstratives, quantifiers, and numerals by corpus

determiner type	omission rate % (marked:not marked)					
	ICE-SIN		ICE-HK		ICE-IND	
demonstrative	3.85	(150:6)	7.87	(164:14)	1.99	(296:6)
no demonstrative	1.41	(3,368:48)	7.56	(3,008:246)	2.31	(3,092:73)
quantifier	2.19	(669:15)	13.20	(651:99)	3.04	(670:21)
no quantifier	1.35	(2,849:39)	6.00	(2,521:161)	2.05	(2,718:58)
numeral	1.98	(842:17)	5.97	(788:50)	2.67	(839:23)
no numeral	1.36	(2,676:37)	8.10	(2,384:210)	2.15	(2,549:56)

The table reads as follows: Among all demonstrative plus noun combinations in ICE-SIN, 3.85 percent of the nouns therein are unmarked. When no demonstrative (but potentially one of the other determiner types) precedes, 1.41 percent of the nouns lack the plural suffix. The nouns can be preceded by more than one determiner type.

Overall, omission rates in both the presence and absence of demonstratives, quantifiers, and numerals are low. Comparatively high omission rates were observed in ICE-HK, particularly when a quantifier is present (13.20 percent). Examples 6.3, 6.4, and 6.5 below contain nouns that lack nominal plural marking preceded by a demonstrative, quantifier, and numeral, respectively.

(6.3) You have also heard from her that Hau Cheun Sum in those *day* before the robbery <,> and a pistol-like object <,> she had seen him loading and unloading the gun (ICE-HK:S2A-066#51:1:A)

(6.4) We said <,> uh lest you waste your breath <,> there are a few *point* on which we will object <,> (ICE-IND:S1B-031#49:1:C)

(6.5) And uh at one time we have got six hundred and sixty *student* taking Statistics in second year (ICE-SIN:S2A-049#74:1:A)

Table 6.5 approaches the data from a different perspective. It focuses on the unmarked nouns and indicates for each corpus the percentage of unmarked nouns preceded and not preceded by a demonstrative, quantifier, and numeral. Thus, instead of comparing the omission rates in the presence and absence of different determiner types, the table shows to which degree unmarked nouns are preceded by one of the determiner types just mentioned. Here, 11.11 percent of the unmarked nouns in ICE-SIN are preceded by a demonstrative and 88.89 percent not. The latter might be preceded by another determiner type though.

Table 6.5: Percentage of unmarked nouns preceded and not preceded by a demonstrative, quantifier, and numeral by corpus (absolute numbers in brackets)

determiner type	ICE-SIN		ICE-HK		ICE-IND	
demonstrative	11.11	(6)	5.38	(14)	7.59	(6)
no demonstrative	88.89	(48)	94.62	(246)	92.41	(73)
in sum	100.00	(54)	100.00	(260)	100.00	(79)
quantifier	27.78	(15)	38.08	(99)	26.58	(21)
no quantifier	72.22	(39)	61.92	(161)	73.42	(58)
in sum	100.00	(54)	100.00	(260)	100.00	(79)
numeral	31.48	(17)	19.23	(50)	29.11	(23)
no numeral	68.52	(37)	80.77	(210)	70.89	(56)
in sum	100.00	(54)	100.00	(260)	100.00	(79)

Above we saw that HKE nouns preceded by a quantifier are comparatively prone to omission, both compared with the other determiner types and the other ICE corpora. By focusing on the unmarked nouns exclusively, we see that of the 260 unmarked nouns in ICE-HK, 38.08 percent (99 nouns in total) are preceded by a quantifier, compared with 5.38 percent (14 nouns) that are preceded by a demonstrative, and 19.23 percent (50 nouns) that are preceded by a numeral. In the SgE

and IndE data, in contrast, slightly higher percentages of the unmarked nouns are preceded by a numeral than by a quantifier. For each combination of corpus and determiner type, the large majority of the unmarked forms lack a preceding determiner of whatever type though. Additionally, only the fewest of the unmarked nouns are preceded by a demonstrative across corpora. In contrast with the low absolute numbers of verbs not marked for past tense preceded or followed by a time adverbial (see section 5.3), we have considerably higher numbers of inflectionally unmarked nouns to work with, particularly in ICE-HK. Still, it is only by means of controlled conditions that the impact of the preceding determiner types on plural omission can be accounted for systematically. The experiment presented in chapter 8 focuses on the perception of unmarked nouns that are preceded by quantifiers because the presence of this determiner type proved to be most influential in ICE-HK. It is worth mentioning in that context that rates of nominal plural omission after quantifiers and numerals are considerably higher than rates of verbal past tense omission after time adverbials. The corpus results on omission of verbal past tense marking showed that time adverbials either play no role for past tense marking or actually "remind" speakers to use the past tense form of the verb. Quantifiers and numerals, in contrast, seem to function as salient plural markers that make the plural suffix redundant.

As pointed out above, *one of* is a particularly noteworthy case because the numeral *one* might trigger a singular form to follow. Luckily, noun plural marking and lack thereof following *one of* can be easily searched for in the ICE corpora, which is why all nouns following *one of* in ICE were manually investigated for lack of plural marking and compared with lack of plural marking after *one of* in the original sample. Table 6.6 provides the omission rates after *one of* both in all nouns preceded by *one of* in the ICE corpora (irrespective of the noun sample) and in the sampled nouns. Only countable nouns following *one of* were considered.

Table 6.6: Omission rates after *one of* among all nouns and the sampled nouns by corpus

corpus	omission rate % (marked:not marked)			
	all nouns		sampled nouns	
ICE-SIN	15.58	(168:31)	13.46	(45:7)
ICE-HK	39.27	(133:86)	45.00	(22:18)
ICE-IND	34.62	(136:72)	29.55	(31:13)

In all three varieties, the majority of the nouns following *one of* are marked for plural, but a considerable percentage lack plural marking in ICE-HK (39.27 percent) and ICE-IND (34.62 percent). Among the sampled nouns, the respective percentages are 45 percent (ICE-HK) and 29.55 percent (ICE-IND). Consequently, the omission rates in all nouns and in the sampled nouns are highly comparable. Consider the following examples:

(6.6) Of course one of my *friend* is <,> really mad after her (ICE-IND:S1A-038#256:1:A)

(6.7) So this is uh one of the *way* for you to narrow it down okay (ICE-HK:S2A-060#150:1:A)

Genre-(un)specificity

Let us have a look at the distribution of plural omission rates across the spoken part of ICE. All spoken sections in ICE were considered to account for usage frequency as a determinant of omission in this small-sized corpus (see section 4.1). Comparable to the results for omission of verbal past tense marking, omission of nominal plural marking in ICE-HK clearly prevails in face-to-face conversations. In ICE-SIN and ICE-IND, omission rates are more evenly distributed and considerably lower than in ICE-HK. The fact that omission in HKE occurs most in spontaneous speech indicates that omission of nominal plural marking most likely slips in when speakers monitor their speech least. The rates of verbal past tense marking in ICE-SIN and ICE-IND were too low to draw respective conclusions.

6.4 A usage-based approach to omission of nominal plural marking

This section approaches nominal plural marking from a usage-based perspective, i.e., it focuses on whether noun frequencies can account for the observed omission rates. Again, the frequencies of use for each of the relevant varieties are approximated by the relative token frequency of each lemma in GloWbE. The relative lemma token frequency was calculated by dividing the absolute lemma token frequency by corpus size to account for the differences in size of the GloWbE corpora. Depending on the status of plural omission as a learner feature or an innovation, omission was assumed to affect infrequent forms more or first and frequent forms less or later (see

Table 6.7: Plural omission rates by corpus section

corpus section	omission rates % (marked:not marked)		distribution of sections %
ICE-SIN:			
Private dialogues	1.83	(1,022:19)	33.33
> face-to-face conversations	2.00	(930:19)	
> phonecalls	0.00	(92:0)	
Public dialogues	0.89	(1,116:10)	26.67
Unscripted monologues	2.93	(696:21)	23.33
Scripted monologues	0.58	(684:4)	16.67
ICE-HK:			
Private dialogues	10.89	(1,318:161)	33.33
> face-to-face conversations	11.13	(1,214:152)	
> phonecalls	7.96	(104:9)	
Public dialogues	5.67	(815:49)	26.67
Unscripted monologues	7.73	(573:48)	23.33
Scripted monologues	0.43	(466:2)	16.67
ICE-IND:			
Private dialogues	2.01	(1,515:31)	33.33
> face-to-face conversations	2.01	(1,413:29)	
> phonecalls	1.92	(102:2)	
Public dialogues	2.86	(781:23)	26.67
Unscripted monologues	2.73	(642:18)	23.33
Scripted monologues	1.53	(450:7)	16.67

hypothesis 1a, section 1.3). This reasoning is in line with the Conserving Effect (cf. Hockett 1958: 180–181, in Bybee 1985: 119; Bybee & Thompson 2007), according to which frequent forms are strongly entrenched in the mind and therefore less prone to change than infrequent forms.

Figure 6.3 plots the omission rates by logarithmically transformed relative lemma token frequency for ICE-SIN, ICE-HK, and ICE-IND. The logarithmic transformation of the relative lemma token frequencies helps visualize the data. Recall that by means of logarithmic transformation the data points are spread more evenly across the graph. Therefore, tight clusters in frequency regions many lemmata fall into are avoided. Each dot depicts a lemma.

As the regression lines in figure 6.3 show, there is a slight downward trend in all three corpora, meaning that omission rates tend to be higher in low frequency lemmata than in high frequency lemmata, which is in line with hypothesis 1a. In

ICE-HK, the omission rates spread widely across the frequency range, which makes it difficult to speak of a clear trend in HKE.

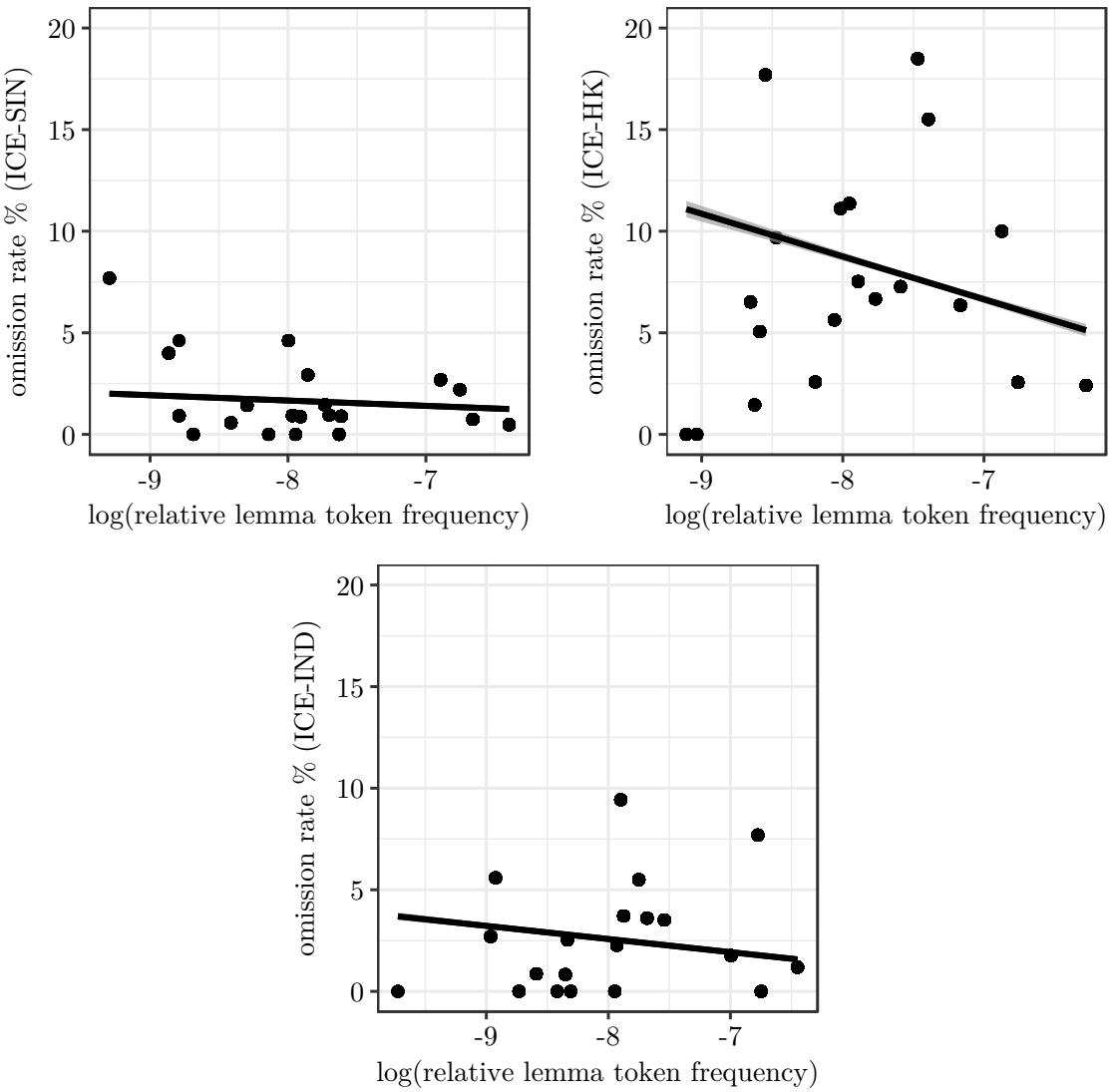

Figure 6.3: Plural omission rates in ICE by relative lemma token frequency, by corpus

When plurality rates are accounted for, the picture changes in ICE-SIN and ICE-HK. The plurality rate of a noun is defined here as the extent to which nouns with a plural reference account for all occurrences of the respective noun in the corpus. It was calculated by dividing the number of occurrences of the lemma in a plural context (marked or unmarked) by the overall number of occurrences of the lemma in ICE (compare table C.1 in appendix C for the respective lemma token frequencies). Figure 6.4 depicts the observed omission rates by plurality rate. Note

that the plurality rates are not logarithmically transformed. Nevertheless, they are comparable to the logarithmically transformed relative lemma token frequencies depicted in figure 6.3. Logarithmic transformation only changes the scale of the x-axis but not the trends indicated by the regression lines.

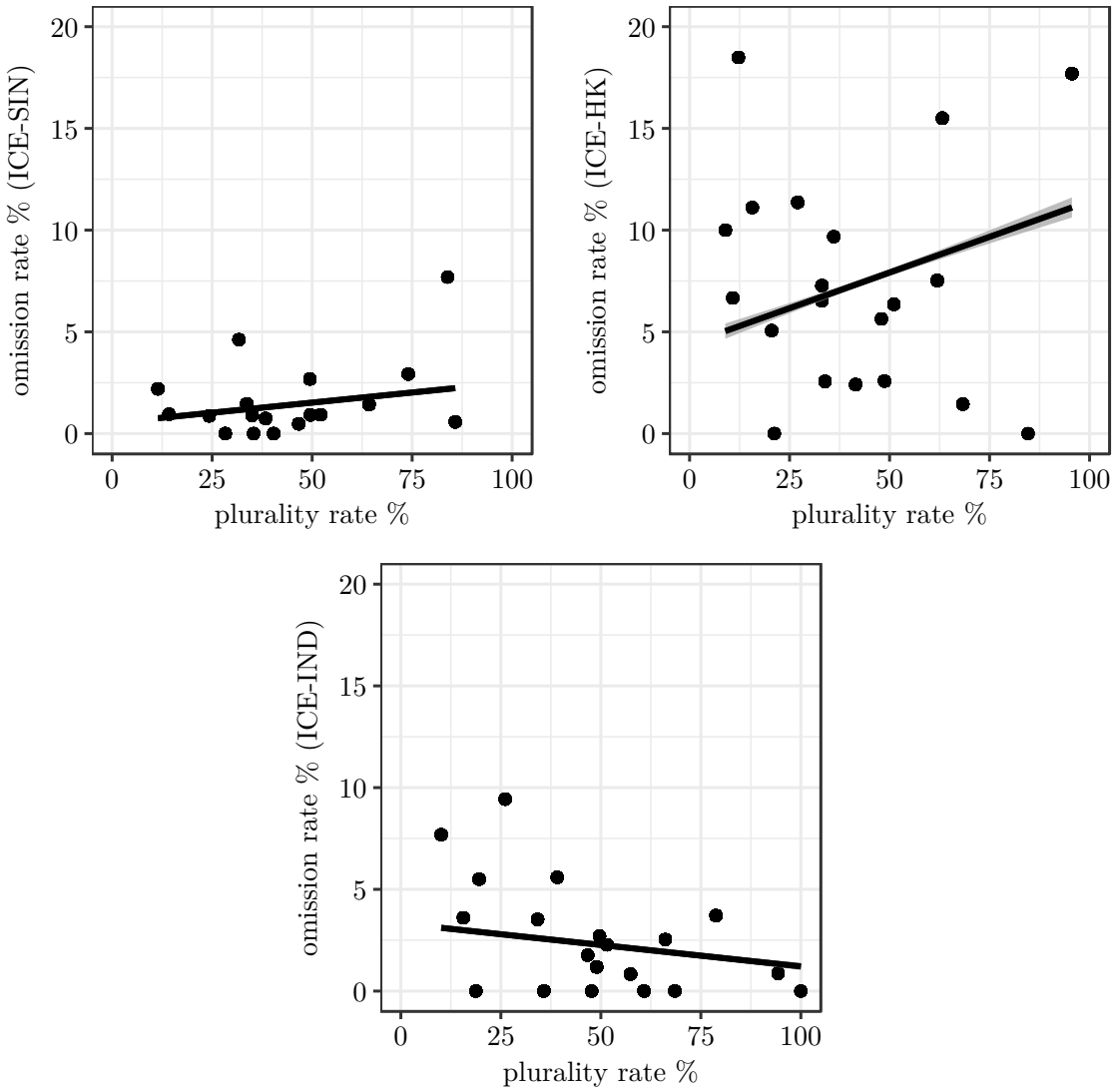

Figure 6.4: Omission rates in ICE by plurality rate, by corpus

Nouns that occur more often in a plural context in ICE-SIN and ICE-HK have comparatively high omission rates. In ICE-HK, no clear picture emerges, and the upward trend is carried by two high frequency outliers, namely *parent* and *student*.[67]

[67] *parent*: plurality rate: 95.59, omission rate: 17.69, marked: 107, not marked: 23; *student*: plurality rate: 63.21, omission rate: 15.51, marked: 376, not marked: 69

It changes to a downward trend once those outliers are discarded. The outlier with comparatively high plurality and omission rates in ICE-SIN is *shoe* (plurality rate: 83.87, omission rate: 7.69, marked: 24, not marked: 2). When this outlier is left out, the overall trend becomes marginal. In ICE-IND, the same downward as trend as for the correlation of omission rates with relative lemma token frequencies is observable; omission rates are comparatively high in infrequent nouns.

To sum up, hypothesis 1a holds when relative lemma token frequencies are considered. I.e., infrequent nouns tend do be affected by omission more than frequent ones. When the plurality rate of a noun is accounted for, the trend changes in ICE-SIN and ICE-HK; in the latter case only when the two outliers are not discarded. In ICE-SIN in particular, nouns that occur often in the plural are comparatively prone to omission. This is counter-intuitive, but as we learned, the upward trend becomes marginal once the outlier *shoe* (with only two unmarked hits, notably) is discarded. Overall, omission rates follow clearer patterns in ICE-SIN and ICE-IND than they do in ICE-HK, which is due to the larger variation in omission rates in the latter corpus.

While Bao's (2010) usage-based approach to substratum transfer is of theoretical relevance both for the corpus studies presented in this book, the model he proposes is not directly applicable. The transfer process for the *give*-passive in SgE was mentioned in section 3.1. *Ho* ("give") is typically used in two morphosyntactic frames in Hokkien: "*ho* NP_1 NP_2," where it functions as a ditransitive verb, and "*ho* NP V," where it is used in passive voice. English provides the morphosyntactic material for the former frame, which is why only this frame can find its way into SgE. One straightforward application of the model to omission of nominal plural marking in SgE would be to test whether preverbal nouns are more affected by plural omission than postverbal nouns. In Chinese, preverbal nouns indicate general reference, whereas postverbal nouns indicate individual reference. Should preverbal nouns be affected by omission more than postverbal nouns, substrate transfer from Chinese is a potential explanation (Zhiming Bao, p.c., 20 September 2016). Regarding omission of nominal plural marking, English provides the morphosyntactic frame to place the regular plural suffix *-s* in or to form an irregular plural form, but when the substrate languages do not mark for plural inflectionally, no respective features can be transferred to the morphosyntactic frame English provides. Many of the IndE substrate languages inflectionally mark for plural themselves anyway (see section 6.5).

6.5 Substratum transfer and institutionalization as constraints on usage frequency

In the previous section, we learned from the ICE-HK data that omission rates differ considerably and without a clear pattern across the frequency ranges the sampled nouns occupy (relative lemma token frequency in GloWbE, plurality rate in ICE). In ICE-SIN and ICE-IND, in contrast, omission rates are low irrespective of relative lemma token frequency and plurality rate. The following paragraphs address the question to which degree substratum transfer and institutionalization explain cross-varietal differences in the omission rates in general and to which degree they serve as constraints on frequency in particular.

Substratum transfer

Section 6.1 revealed that lack of nominal plural marking has been described for both SgE and HKE. Its equivalence in form with singular countable nouns that are used without articles is particularly remarkable. While the literature on IndE has addressed article omission, omission of nominal plural marking has not been an issue. From the corpus analyses, we learned that omission is comparatively frequent in ICE-HK and infrequent in both ICE-SIN and ICE-IND (section 6.3).

One likely reason for the observed omission rates in ICE-HK is that in the substrate Cantonese pre-nominal elements like numerals "serve to indicate to the speaker that the following noun has plural reference" (Budge 1989: 44); not noun inflection. Consequently, speakers of HKE might omit the plural suffix because Cantonese does not inflectionally mark its nouns for plural. Since nouns along the frequency cline are affected to similar degrees, substratum transfer likely constrains potential frequency effects (compare hypothesis 2, section 1.3). Interestingly, the various Chinese substrates of SgE do not inflectionally mark nouns for plural either, and plural omission rates in ICE-SIN are vanishingly small. Recall Biewer's (2015: 172) argument that perceptually salient markers for plurality (e.g., quantifiers and numerals) are likely to be preferred over less salient markers (the plural suffix). Early learners in particular opt for salient markers at the expense of seemingly redundant ones. Consequently, we could argue that speakers of HKE, who are more learner-like than speakers of SgE, are more likely to choose salient markers such as quantifiers or numerals as a surrogate for the plural suffix than speakers of SgE. Table 6.4, in fact, revealed that omission rates after demonstratives, quantifiers, and numerals

are higher in ICE-HK than they are in ICE-SIN. However, it has to be pointed out that the same is true when no determiner precedes the sampled nouns.

In contrast with the corpus study on omission of verbal past tense marking, which hardly revealed any omission in the IndE data, omission rates of the regular plural suffix in IndE are well comparable to those in the HKE data; despite the fact that Indian languages such as Hindi, Tamil, and Kannada have inflectional plural markers (for Hindi, see Y. Kachru 2006: 44–45; for Tamil, see Lehmann 1993: 11–13; for Kannada, see Schiffman 1983: 23). Let us have a look at the language background of the speakers who omit nominal plural marking in ICE-IND. Table 6.8 depicts information on the age, gender, birthplace, mother tongue, level of education, and overseas experience the respective speakers in ICE-IND and ICE-HK.

As with the background information of speakers who omit inflectional past tense marking, the homogeneity of the speaker population in ICE-HK that omits nominal plural marking is worth stressing. The large majority of the unmarked nouns in ICE-HK were produced by speakers who were born in Hong Kong and who speak Cantonese as their mother tongue. Of the speakers who were not born in Hong Kong and who have a mother tongue other than Cantonese, all except for five speakers (for whom no respective information is available) received primary and/or secondary education in Hong Kong. 111 of the 260 unmarked nouns were uttered by speakers aged between 21 and 25. While no age cline comparable to that among the speakers who omit verbal past tense marking is observable (younger speakers omitted verbal past tense marking more than older speakers), the fact that many of the unmarked nouns were uttered by young speakers can be interpreted as a sign of emerging conventionalization.

As mentioned above, the ICE-IND speakers are a very heterogeneous group, which is particularly obvious from the variety of birthplaces and mother tongues listed. The Indian mother tongues stem from two different languages families, comprising Dravidian languages (Kannada, Tamil, Malayalam, and Telugu) and Indo-Aryan languages (Marathi, Hindi, Bhojpuri, Konkani, Sindhi, and Kashmiri; compare table 3.5 in section 3.4). As we learned above, Indian languages like Kannada, Hindi, and Tamil, have inflectional plural markers, so substrate influence cannot explain why they should omit the regular plural suffix *-s*. The various speaker backgrounds underline the impression that omission of nominal plural marking occurs sporadically in IndE and is not tied to specific speaker groups. However, we will see in chapter 8.1 that participants from India who took part in the web-based experiment

Table 6.8: Background of speakers who omit inflectional marking for plural

	ICE-HK (260†)	ICE-IND (79†)
age	21–25: 111, 36–40: 32, 20 and below: 30, 41–45: 21, 56–60: 12, 51–55: 12, 46–50: 12, 26–30: 11, 31–35: 9, above 60: 6, no info: 4	above 50: 26, no info: 18, 18–25: 14, 24–33: 13, 34–41: 5, 42–49: 2, above 40: 1
gender	female: 139, male: 117, no info: 4	female: 20, male: 57, no info: 2
birthplace	Hong Kong: 216, China: 33, no info: 8, United Kingdom: 1, Macao: 1, South Korea: 1	no info: 53, Karnataka (state): 17, Maharashtra (state): 3, Bombay: 2, Gambat(Sind), Pakistan: 1, Kolhapur: 1, Uttar Pradesh (state): 1, Andhra Pradesh (state): 1
mother tongue	Cantonese: 245, Chinese: 9, no info: 4, Mandarin: 1, Korean: 1	no info: 21, Kannada: 17, Marathi: 16, Hindi: 6, Tamil: 5, English: 3, Bhojpuri: 2, Konkani: 2, Malayalam: 2, no info: 2, Sindhi: 1 Kashmiri: 1, Telugu: 1
level of education	University: 141, Secondary: 101, no info: 11, Primary: 7	no info: 43, Master's degree: 20, Graduate: 11, Doctorate: 4, Undergraduate: 1
overseas experience	37 (no info: 223)	0 (no info: 79)

†Number of nouns lacking nominal plural marking. Speakers may be represented multiple times per cell depending on the number of unmarked nouns they produced.

read inflectionally unmarked nouns without a preceding quantifier and the words directly following them comparatively fast and evaluated them comparatively positively. This is one of the stimuli they were presented with: *Bill told us detail about the terrible accident he had been involved in* (see table 8.3 in section 8.2.2). The finding points towards familiarity with and approval of bare noun phrases in this group and might explain why omission of nominal plural marking occurs in IndE despite the fact that relevant substrate languages mark plural inflectionally.

Institutionalization

Let as elaborate on the question whether institutionalization functions as a constraint on frequency effects in the last step. This was assumed to be the case should patterns of plural omission be particularly stable in SgE, the most institutionalized variety considered, irrespective of lemma token frequency (see hypothesis 3, section 1.3). SgE is in stage 4 of Schneider's (2003; 2007) Dynamic Model, whereas HKE and IndE are in stage 3. According to Ooi (2001a), the increasing conventionalization of variety-specific features in SgE even justifies to speak of a transition of the variety to stage 5. Similarly, Mukherjee (2007) argues that IndE is on its way to stage 4 because lasting protests against Hindi in the South have led to identity reconstructions that are typical of that stage.

As with omission of verbal past tense marking, omission of nominal plural marking is comparatively frequent in ICE-HK, but omission rates do not exceed 20 percent across varieties.[68] The very low omission rates in ICE-SIN and ICE-IND justify to speak of a by chance lack of the plural suffix in those varieties instead. But how stable are the observed omission patterns in the varieties of interest? Omission rates tend to decrease in all three contact varieties with increasing lemma token frequencies and increasing plurality rates (except for a slight upward trend due to the outlier *shoe* in ICE-SIN). Compared with ICE-SG and ICE-HK, omission along the frequency cline is comparatively little patterned in ICE-HK.

Regarding preceding demonstratives, quantifiers, and numerals, omission rates in nouns with premodifiers are higher in ICE-HK than in ICE-SIN and ICE-IND, but a clear tendency for omission is only visible after quantifiers; and only in ICE-HK (tables 6.4 and 6.5). Substratum transfer is likely an explanation (see above). According to Biewer (2015), the fact that (early) learners prefer cognitively salient options is a sign of "an application of Slobin's *Principle of economy of production*

[68] Recall that rates of verbal past tense omission reach 30 percent in ICE-HK, and particularly so in cases where the *-ed* suffix is prone to consonant cluster reduction.

in an L2 setting (Slobin 1973; 1977; Williams 1987: 169)" (173). Additionally, she points out that "L2 learners in general have difficulties with bound morphemes at the beginning of the learning process and tend to leave them out because of these difficulties (Winford 2003: 218)" (Biewer 2015: 173). This explains why plurality tends to be marked by means of a quantifier rather than by inflection in HKE. The described effect is probably enhanced by the unfamiliarity of speakers of HKE with inflectional marking from Cantonese.

On the basis of Van Rooy's (2011) account of learner errors versus conventionalized innovations in New Varieties (see section 2.4), the following can be deduced: Omission of nominal plural marking in HKE constitutes a learner error rather than a conventionalized innovation because it does not occur systematically and is likely due to transfer from Cantonese. The low and unsystematic rates of plural omission in ICE-SIN und ICE-IND imply that lack of nominal plural marking occurs by chance in SgE and IndE. Substrate influence cannot explain plural omission in IndE, but it can account for the comparatively few instances of omission of the plural suffix in SgE.

6.6 Concluding remarks

As in the corpus study on omission of verbal past tense marking, all three determinants of simplification considered (usage frequency, substratum transfer, and institutionalization) contribute to explaining the observed plural omission rates. Omission rates turned out to be unexpectedly low and again the question emerges whether plural omission is a salient rather than a particularly frequent feature in the Asian contact varieties of interest. One likely explanation of the low degrees of plural omission observed is the speaker group recorded for the ICE corpora (section 6.5). The ICE-HK and ICE-IND components mainly comprise highly educated speakers, and many of the face-to-face conversations (defining the most informal part of ICE) took place at university. Respective metadata are lacking for ICE-SIN.

7 Regularization: A corpus-based account

Another simplification process investigated here is regularization. While omission of inflectional marking for verbs and nouns describes simplification on the structural level (compare chapters 5 and 6), regularization is an example of system-based simplification. Regularization of irregular verbs (example 7.1) and the marking of uncountable nouns for plural (example 7.2) are the phenomena of interest here.

(7.1) So I *catched* up on the lost hours of sleep and had loads of dreamy time (GloWbE IN)

(7.2) Apart from that, any *advices* in terms of what else to bring and should do/visit in the park? Much thanks for all the *advices* (GloWbE SG)

Irregular verbs that are regularized take over the past tense formation strategy adopted by the majority of verbs, namely /t,d/ affixation. Similarly, regularized uncountable nouns take the plural suffix *-s* attached to the majority of nouns that refer to plural entities. As with the corpus studies on omission of verbal past tense and nominal plural marking, the impact of usage frequency, substratum transfer, and institutionalization on regularization rates is of particular interest. Details follow in sections 7.1 and 7.2.

A first look at ICE revealed that the corpus is much too small to investigate the regularization phenomena of interest therein. To the author's knowledge, GloWbE is the only available corpus of sufficient size that provides directly comparable data sets for the varieties of interest. Obviously, investigating regularization in GloWbE shifts the focus from spoken language to internet language, which is why the corpus studies presented in the following sections provide insights into language use on the web rather than on regularization in spoken language. Compare section 4.2 on using the web as a resource in corpus linguistics.

Table 7.1 provides the ratings for the features of interest in eWAVE. Both features have B-ratings in CSE and A-ratings in HKE, but only feature 55 (different count/mass noun distinctions resulting in use of plural for StE singular) has been

Table 7.1: The regularization features in eWAVE (Kortmann & Lunkenheimer 2013)

feature no.	ratings†			top L2	top Asia	L2 simple
	CSE	HKE	IndE			
55	B	A	A	✓	✓	
128	B	A	C			✓

feature no.	feature
55	different count/mass noun distinctions: use of plural for StE singular
128	regularization of irregular verb paradigms

†ratings in eWAVE: A - feature is pervasive or obligatory; B - feature is neither pervasive nor extremely rare; C - feature exists, but is extremely rare; D - attested absence of feature; X - feature is not applicable; ? - no information on feature is available

attested for IndE. Feature 128 has been described as an L2-simple feature, whereas feature 55 is both a top L2 feature and a top Asian feature in eWAVE.

Section 7.1 deals with the regularization of irregular verbs in the Asian varieties of interest, providing both insights into previous research on the phenomenon and embedding the results from the GloWbE searches therein. Section 7.2 investigates the use of uncountable nouns as countable nouns. Section 7.3 embeds the findings for both phenomena into the larger picture.

7.1 Regularization of irregular verbs

As mentioned before, the regularization of irregular verbs has been an ongoing process in the history of English, as the majority of verbs English have undergone regularization and form their past tense by means of /t,d/ affixation today. Verbs that have remained irregular up until now used to be or still are highly frequent in use (Conserving Effect; e.g., Bybee 2007: 10). They have not been prone to regularization because their frequency of use has strengthened their memory representation, which eases their retrieval from the mind (ibid.). Croft & Cruse (2004) additionally argue that the past tense forms of irregular verbs are "listed in the lexicon" (292) as such, whereas those of regular verbs are retrieved by adding the highly productive *-ed* suffix to the verb base.

The regularization of irregular verbs is cognitively costly. Nevertheless, the feature has been categorized as an L2-simple feature by Mesthrie (2012), as table 7.1 above

shows. For learners of English, the acquisition of irregular verbs is a matter of learning effort, and the marking of the verb by means of the regular *-ed* suffix is likely a sign that the irregular past tense form of the verb has not been acquired or is not remembered. In eWAVE, the feature is only rated as pervasive or obligatory for HKE, which speaks for the learner argument just described. The next section presents previous research on the regularization of irregular verbs in the contact varieties of interest, including information on potential substratum transfer.

7.1.1 State of the art

Singapore English

The regularization of irregular verb forms is hardly an issue in the literature on SgE, although eWAVE lists the regularization of irregular verb paradigms as "neither pervasive nor extremely rare" in CSE (B-rating). Obviously, its high salience is likely to account for this feature rating. One exception is Sheng (2007), who investigates past tense marking in both regular and irregular verbs in the spoken part of ICE-SIN. Sheng (2007: 281–284) detects different patterns of use of irregular verbs, and one of them is the use of the regular *-ed* suffix with irregular verbs. In fact, the regularization of irregular verbs occurs very rarely, with *weaved* (occurring twice), *binded*, *winded*, and *leaded* (each occurring once) being the only hits identified. Compare the following examples provided by Sheng (2007: 282; emphases in the original):

(7.3) There's another intricate pattern being *weaved* as the band also play (ICE-SIN:S2A-006#5:1:A)

(7.4) Irene will also show you the correct grammar and the bit stream base lift which is also *binded* in our program (ICE-SIN:S2A-055#5:1:A)

Further patterns detected are the use of the participle instead of the simple past (e.g., *He seen and came out and tell me*, ICE-SIN:S1A-067#107:1:B) and the lack of past tense marking with irregular verbs (e.g., *And in some instances she didn't exactly give yes. In fact sometimes she give no*, ICE-SIN:S1B-064#82-83:1:B). Although such patterns occur only sporadically as well, they are far more frequent than the regularization of irregular verb forms. None of the usage patterns are speaker-specific. These are clear indicators that irregular verb forms are strongly entrenched in the mind. While the participle is sometimes used instead of the simple past or

past tense marking is lacking completely, a change of the simple past and participle forms of irregular verbs is highly unlikely.

According to Davydova (2011), the use of lone participles (which Sheng 2007 terms the use of the participle instead of the simple past) is a result of "universal strategies of simplification frequently employed in second-language development" (246). The feature also occurs in IndE, IndSAfE, St. Helena English, and Butler English (e.g., Mesthrie 2008; Wilson & Mesthrie 2004; Hosali 2008). This finding goes in line with the comparatively frequent use of participles instead of simple past forms in Sheng's (2007) analysis. No similar observations have been made for the use of the *-ed* suffix with irregular verbs, to the author's knowledge. Worth noting in that context is Chamber's (2004: 12) suggestion that the leveling of irregular verb forms constitutes an example of a vernacular universal, i.e., a non-standard feature that occurs across varieties of English. Two examples he provides are *Yesterday John seen the eclipse* and *Mary heared the good news*, which are, strictly speaking, two different phenomena. While *heared* is an example of regularization (however, in the sense that the vowel is not changed), *seen* is a participle that is used instead of the simple past form. The latter has been observed in various varieties of English (see above), the former has been little reported.

Hong Kong English

Despite the fact that over-regularization of various types is a typical phenomenon in (second) language acquisition (e.g., Kirkici 2010), the feature is not elaborated on in the literature on HKE. At the same time, the regularization of irregular verb paradigms is rated as "pervasive or obligatory" (A-rating) in eWAVE (see table 7.1). This allows for two interpretations. Either, the feature is rare in HKE, but when it occurs it is highly salient, or it occurs with a certain frequency but has not (yet) received scholarly attention.

Kirkici (2010: 76–77) investigates L1 Turkish learners of English as an L2 and observes that both the tested advanced learners of English and the tested learners with low-level proficiency in English regularized comparatively man infrequent irregular verbs. Furthermore, the L2 subjects with low proficiency in English "resorted more often to the regularization of LF [low frequency] irregular forms" (ibid.: 77) than the L2 speakers with advanced proficiency in English. Even the L1 control group regularized 3.57 percent of the low frequency irregular verbs it was presented with. The results clearly indicate that the past tense forms of low frequency irregular verbs are less remembered than those of high frequency irregular verbs; particularly so among

early learners of English. It should be pointed out, however, that Turkish learners acquire English as a foreign language, whereas HKE is spoken in a setting where English is one of the official languages[69]. The fact that even L1 speakers regularize some low frequency irregular verbs shows that the regular *-ed* suffix is attached to the stem of infrequently used verbs across speaker types.

Indian English

Davydova (2011: 181), who has been referred to above with regard to the regularization of irregular verbs in SgE, also elaborates on patterns of past tense marking in IndE. In contrast with ICE-SIN, where at least a few instances of regularized irregular verbs were found, she reports no respective hits for ICE-IND. As in ICE-SIN, she observes single instances of lone participles that occur in place of simple past forms though. Since the lone participles occur in present perfect contexts only, Davydova (2011) argues that "these forms seem to be constructions in which the auxiliary *have* has undergone deletion as a result of imperfect L2 acquisition" (181). As mentioned above, such "universal strategies of simplification [are] frequently employed in second-language development" (ibid.: 246) and have been found in IndSAfE, St. Helena English, and Butler English as well. The lack of regularized irregular verbs is a clear sign of the strong entrenchment of irregular past tense forms in the mind instead.

7.1.2 Corpus findings

All irregular verbs were considered that occur with a lemma token frequency of at least 1,000 in GloWbE. The verb lemma token frequency lists for each variety retrieved from the GloWbE offline database for the corpus study on omission of verbal past tense marking (chapter 5) were used here as well. The regularized verbs were retrieved by means of the search mask that comes with the online version of GloWbE.[70] The number of hits was vanishingly small across lemmata and allowed for manual analyses of the hits obtained, which made it possible to identify and discard wrongly tagged forms. GloWbE GB served as the control corpus.

Table 7.2 summarizes the hits for irregular verbs that adopt the regular past tense *-ed* suffix ("reg."). It only contains lemmata which occur five times or more in any of

[69] Nevertheless, Cantonese is the preferred choice in the private domain, which is why the status of HKE as a learner variety is a matter of debate (see section 3.3.1).

[70] The search syntax for *costed*, for instance, was "costed.[v*]." All instances of formal past tense marking (compare section 5.3) were considered.

the corpora considered. Even five hits are vanishingly small for a large corpus such as GloWbE (and compared with the hits for standard-like irregular occurrences of those lemmata), but as the table reveals larger numbers of regularized verb forms are a clear exception. GloWbE IN is roughly twice the size of GloWbE HK and GloWbE SG, which is visible in the lemma frequencies and needs to be kept in mind when reading the table. The regularization rate ("r(eg). rate") was calculated by dividing the number of verbs ending in the regular past tense *-ed* suffix by the sum of those hits and the hits for verbs that adopt the StE irregular past tense marking ("irreg.")[71] and is provided as a percentage. The absolute lemma token frequency of each lemma per corpus is also listed ("l(em). freq."). The relative lemma token frequencies are too small to meaningfully provide them in the table. For ease of interpretation, the entries are sorted by the sum of regularized verbs in all three GloWbE corpora of interest ("sum reg."). Finally, the morphological process that underlies the StE past tense form of the respective lemma is given ("morph. process"). Grammatically wrong forms were left in the analyses (e.g., *I didn't heared before of some people ...*, GloWbE GB), whereas proper names were excluded. Occurrences of *leaded* as an adjective (as in *leaded patrol* or *leaded window*) and past tense *puttet* (as in *I played really poorly the first couple of days and putted great*, GloWbE IN) were not considered either.

Notably, hardly any instances of a regularization of irregular past tense verb forms are observable in GloWbE, i.e., the regularization rates are very small for all three contact varieties and the control variety. Since GloWbE represents written language (see section 4.1 on internet language), the language data therein are more prone to reflection than spoken language would be. This potentially contributes to the small regularization rates detected. In contrast with the eWAVE ratings, according to which the regularization of irregular verbs is a salient feature in HKE, no differences are observable across the corpora. The morphological process underlying the irregular past tense formation of the lemmata listed does not impact on degrees of regularization either. Additionally, for none of the corpora the regularization rates increase or decrease with higher lemma token frequency. Lemmata like *lead*, *choose*, *hear*, or *put* are comparatively frequent throughout, but they are not more or less prone to regularization than the less frequent lemmata investigated. *Heared* and *layed*, finally, might simply constitute misspellings of *heard* and *laid*.

[71] The part-of-speech tag used for unique word forms such as *arose* was ".[v*]." The past tense and past participle forms of verbs that form their past tense by means of zero marking were identified by means of the tags ".[vvd]" and ".[vvn]."

Table 7.2: Regularization of irregular verb forms in GloWbE (five occurrences or more in any corpus, sorted by sum reg., r(eg). rate in %)

lemma	GloWbE SG				GloWbE HK				GloWbE IN				sum	GloWbE GB		morph.
	reg.	irreg.	r. rate	lem. freq.	reg.	irreg.	r. rate	l. freq.	reg.	irreg.	r. rate	l. freq.	reg.	reg.	r. rate	process[†]
cost	21	330	5.98	12,357	14	221	5.96	11,403	29	599	4.62	25,423	64	235	5.35	ZM
lead	4	3,834	0.10	3,821	10	4,379	0.23	2,655	9	11,146	0.08	5,911	23	11	0.02	VC
arise	4	289	1.37	21,316	4	422	0.94	17,743	15	1,199	1.24	42,737	23	7	0.19	VC
stick	10	1,604	0.62	12,473	5	1,014	0.49	10,057	7	2,425	0.29	22,169	22	15	0.10	VC
choose	3	3,733	0.08	11,958	4	3,213	0.12	12,264	14	8,067	0.17	31,395	21	21	0.06	VC
seek	9	978	0.91	5,661	1	1,161	0.09	6,122	10	4,180	0.24	15,901	20	40	0.37	VC & A
hear	1	6,114	0.02	4,306	1	4,960	0.02	3,260	17	11,062	0.15	6,765	19	52	0.08	VC & A
hurt	3	274	1.08	2,583	2	212	0.93	2,429	14	468	2.90	7,200	19	6	0.11	ZM
lay	5	979	0.51	2,046	2	1,195	0.17	1,743	10	3,643	0.27	4,918	17	135	1.14	A
strike	9	854	1.04	6,609	1	772	0.13	3,992	6	2,331	0.26	10,071	16	18	0.15	VC
put	0	17,748	0.00	2,137	5	15,308	0.03	2,880	8	36,388	0.02	6,277	13	3	0.00	ZM
catch	4	2,911	0.14	2,657	3	1,903	0.16	1,643	6	5,513	0.11	5,056	13	26	0.11	VC & A
spread	4	873	0.46	2,527	1	1,023	0.10	2,825	5	2,850	0.18	7,409	10	12	0.15	ZM
wake	2	110	1.79	2,926	0	84	0.00	1,668	8	264	2.94	4,421	10	51	0.88	VC

[†]A - affixation, VC - vowel change, ZM - zero marking

As regards the impact of lemma token frequency on omission rates, it was hypothesized that infrequent irregular forms are affected by regularization more (or first) and frequent irregular forms less (or later; see hypothesis 1b, section 1.3).

In figure 7.1, the regularization rates for all irregular verbs (i.e., not only those depicted in table 7.2 that occur at least five times) investigated are plotted against the logarithmically transformed lemma token frequencies of the respective verbs. While an investigation of the regularization rates by lemma token frequency for the most frequently regularized verbs listed in table 7.2 revealed no clear trends, considering all verbs investigated shows the following: In all the corpora, regularization rates tend to be higher for verbs with smaller lemma token frequencies. This is the case for GloWbE SG and GloWbE HK in particular and goes in line with the hypothesis. As mentioned above, Kirkici's (2010: 77) study revealed that even L1 English speakers regularize a certain percentage (3.57 percent) of low frequency irregular verbs, which is a clear sign of a tendency for regularization when the irregular past tense form cannot be recalled. Compared with the eWAVE ratings for HKE ("pervasive or obligatory") and CSE ("neither pervasive nor extremely rare"), the GloWbE data paint a much more conservative picture, but we have to keep in mind that GloWbE consists of internet language, whose production is more monitored than speech.

Returning to table 7.2, the regularization rate of *cost* turns out to be considerably higher than that of the other irregular verbs analyzed. It is likely that analogy plays a role, i.e., as with regular verbs the *-ed* suffix is attached to the base form, resulting in *costed*. Here are two examples:

(7.5) A pair of sunglasses originally *costed* $250 for one pair (GloWbE HK)

(7.6) Everyone raved about it and the cauliflower *costed* me only $1.30. Hoo-rah! (GloWbE SG)

The other three lemmata in the list that do not overtly mark past tense either (*hurt*, *put*, and *spread*) do not differ in their regularization rates from the irregular verbs that mark their past tense by means of a morphological process other than zero marking. Three examples where regularization occurs are:

(7.7) If I *hurted* you or troubled you I am extremely sorry for that (GloWbE IN)

(7.8) i took out the batter and *putted* back inside (GloWbE IN)

(7.9) We sliced the cake into half and then *spreaded* an even layer of cream on it ... (GloWbE SG)

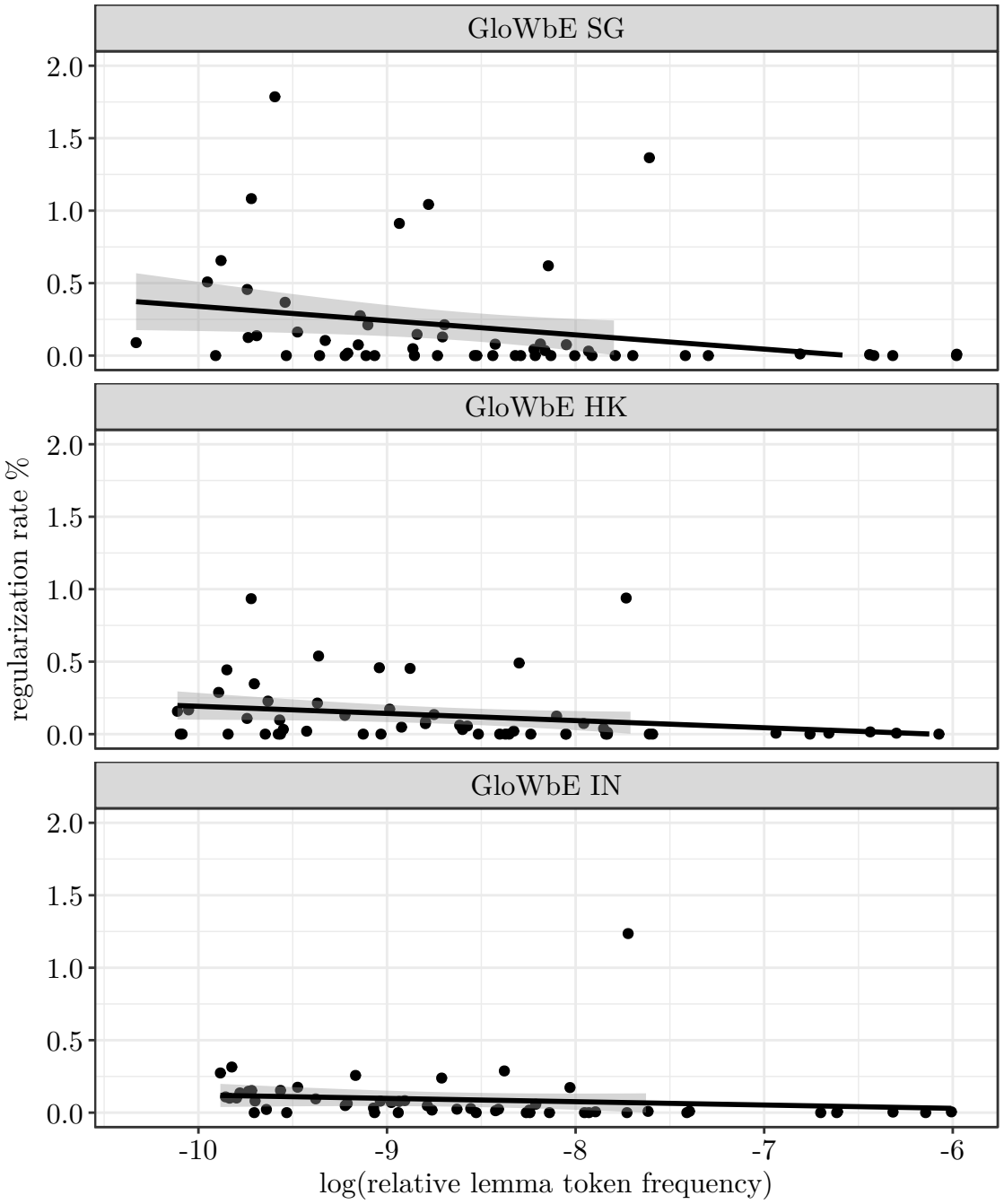

Figure 7.1: Regularization rates in GloWbE by log(relative lemma token frequency), by corpus

To sum up, regularization occurs too sporadically to be called an established feature in any of the three varieties. The eWAVE ratings for CSE (B-rating) and HKE (A-rating) are likely explicable by the fact that regularized irregular verbs are particularly salient. The overall impression gained is that the irregular past tense forms are too entrenched in the mind to be regularized by any of the speaker groups of interest.

7.2 Uncountable nouns treated as countable nouns

Let us turn to uncountable nouns that attach the highly productive plural suffix *-s* next. According to Biber et al. (2007), "[u]ncountable nouns refer to entities which cannot be counted and do not vary for number" (241). Countable nouns, in contrast "refer to entities which can be counted; they have both singular and plural forms (ibid.). Table 7.3 provides an overview of contrasts between singular and plural as well as indefinite and definite forms of the noun types just mentioned. Both countable and uncountable nouns occur in definite and indefinite uses.

Table 7.3: Main types of nouns (adopted from Biber et al. 2007: 241)

	common countable		common uncountable		proper
	indefinite	definite	indefinite	definite	
singular	*a cow*	*the cow*	*milk*	*the milk*	*Sue*
plural	*cows*	*the cows*	—	—	—

A closer look at individual lemmata reveals that "[t]he use of a noun as countable or uncountable is lexically restricted, and the difference in meaning varies to a great extent with the individual noun" (ibid.: 243). Besides typical uncountable nouns like *milk*, Biber et al. (2007) distinguish, among other things, between zero plurals, plural-only nouns, and collectives. All three noun types are worth zooming into here to see how they relate to uncountable nouns.

Zero plurals are nouns that "have no overt plural ending, although they have plural meaning and concord" (ibid.: 288). Examples Biber et al. (2007: 288–289) mention are words for animals like *duck* and *fish*, quantifying nouns like *pound* (BrE), and nouns whose stem ends in *-s* like *means* or *series*. Since zero plurals refer to countable entities, they are no example of uncountable nouns. However, this characteristic makes them particularly likely to take the plural suffix *-s*.

Plural-only nouns, in contrast, either "do not have a singular-plural contrast" (ibid.: 289), as in *scissors* which only occurs in the plural (**scissor*), or their corresponding singular form differs in meaning from the plural form (e.g., *custom*, which is countable, versus *customs*, meaning "customary behavior"). *People* takes plural concord and means "nation, tribe, race" (ibid.), when it is used as a regular countable noun with both singular and plural concord. Nouns like *news*, *darts* (as in *There's no darts tomorrow*), or *measles*, which "look like plurals but actually behave like uncountable singular nouns" (ibid.: 290) are further examples of plural-only nouns. Nouns ending in *-ics* are usually treated as singular forms when academic disciplines are meant. When those nouns occur with plural concord, they "refer to specific instances of economic facts" (ibid.: 182), as in *Detailed statistics are not available for the inner city itself*. In any case, they keep their plural-only form. In general, it can be assumed that nouns which have a corresponding (and not too infrequent) singular form which differs in meaning from the plural are particularly likely to adopt the plural suffix (compare *custom*, *people*).

The last group are collective nouns, which "refer to groups of single entities" (ibid.: 247). Since they occur both in the singular and in the plural as well as in indefinite and definite uses, those nouns do not differ in their behavior from countable nouns and are clearly no example of uncountable nouns. Examples Biber et al. (2007) mention are *army*, *audience*, *board*, *committee*, *crew*, *family*, *jury*, *staff*, and *team*. *Staff* is a special case in that it mainly occurs with plural concord (ibid.: 188).

Table 7.4 provides an overview of the different noun types just presented and additionally lists nouns with regular and irregular plural marking. For each noun type, the table summarizes whether a formal singular-plural contrasts exists, whether uses of nouns in the singular differ in meaning from uses in the plural, and whether nouns of the respective noun type take singular or plural concord. The most straightforward cases are nouns with regular and irregular plural marking. They formally distinguish between singular and plural, take singular and plural concord, and their singular forms do not differ in meaning from the plural forms, with the exception that they refer to singular versus plural entities, respectively. Zero plurals (e.g., *fish, pound*), in contrast, lack overt plural marking, but they have plural meaning and plural concord. Plural-only nouns are tricky because they have no singular-plural contrast by definition (compare *scissors*), but members of the group like *custom*, *people* and *news* are special cases (see above). They account for potential formal singular-plural contrasts, meaning differences between singular and plural, and sin-

gular concord (indicated by the bracketed check marks in table 7.4). Collectives behave similarly to nouns with regular and irregular plural marking. While singular concord is the preferred choice (in BrE), a focus on the group rather than on individual members justifies plural concord. Uncountable nouns, finally, only occur with singular concord and have no formal singular plural contrast, which also excludes potential differences in meaning between singular and plural forms.

Table 7.4: Noun type characteristics

noun type	formal sg. pl. contrast	meaning difference sg. pl.	singular concord	plural concord
countable nouns with reg. plural marking	✓		✓	✓
countable nouns with irreg. plural marking	✓		✓	✓
zero plurals				✓
plural-only nouns	(✓)	(✓)	(✓)	✓
collectives	✓		✓	✓
uncountable nouns			✓	

Only uncountable nouns which take the regular plural suffix are of interest here. As we have seen above, zero plurals, plural-only nouns, and collectives all differ from uncountable nouns in various respects, and while a comparison of non-standard plural marking in those noun types with that in uncountable nouns would be worth conducting, it goes beyond the scope of this study. Quirk et al. (1985: 246) point out that uncountable nouns do not only lack pluralization but are additionally not accompanied by an indefinite article in standard usage (see section 7.2.1). Since it is difficult to account for articles preceding or not preceding uncountable nouns in a large corpus such as GloWbE, this aspect will not be considered here. The next section introduces previous research on the use of the plural suffix with uncountable nouns in the contact varieties of interest, including insights on potential substratum transfer.

7.2.1 State of the art

Singapore English

In contrast with the regularization of irregular verb forms, which is hardly an issue in the literature on SgE, the "treat[ment of] non-count nouns as count" (Wee & Ansaldo 2004: 63) has been reported by various authors. Wee & Ansaldo (2004: 63) account for both the use of non-count nouns as count nouns and the opposite pattern, namely the use of bare nouns instead of marked count nouns (e.g., *She queue up very long to buy ticket for us*, taken from Alsagoff & C. L. Ho 1998: 143, in Wee & Ansaldo 2004: 63). The authors suggest that *ticket* "may in fact be used as a non-count or mass noun in such instances" (ibid.: 64) because the noun occurs either uninflected without premodification or inflected with a preceding quantifier. Wee & Ansaldo (2004: 64) present examples from the GSSEC as counterevidence against the latter claim (compare section 6.1). In line with Gil (2013), the authors additionally suggest that "number marking in CSE is essentially sporadic or optional" (ibid.: 64). On the basis of a questionnaire on basilectal SgE, Gil (2013) argues that "bare, unmarked nouns can be interpreted as either plural or singular" (Wee & Ansaldo 2004: 65), following substrate languages like Cantonese, Malay, and Hokkien (cf. Platt & Weber 1980, in Wee & Ansaldo 2004: 65).

Ziegeler (2015) points out that in contrast with StE, where bare nouns always represent uncountable nouns, CSE (or "Singapore Colloquial English," as she refers to it) allows for unmarked countable nouns as well. Instead of claiming that countable nouns are treated as uncountable nouns in Singlish, she argues that there is not necessarily a countability requirement for unquantified noun phrases in CSE because the Chinese substrate languages of SgE do not have the countability requirement StE has (ibid.: 183). In contrast, Liu et al. (2006), who investigate the language data of Chinese learners of English, stress that while "[i]n a broad sense, the terms count and non-count nouns are conceptualized in the same way in English and Chinese [. . .] differences exist in how individual lexical items are categorized" (136). The authors specify that most nouns that are considered count in Chinese are premodified by a classifier (e.g., *san zhang yizi*, "three + classifier + chair, i.e., three chairs") and those considered non-count follow a measure word (*liang bei kafei*, "two + measure word + coffee, i.e., two cups of coffee"). The lack of such a consistent distinction in English "can be a source of confusion for Chinese learners" (ibid.: 137). Abstract nouns like *desire*, *attitude*, or *thought*, which are context-dependently used as count

or non-count nouns in English, are specifically mentioned (Liu et al. 2006: 137). This also explains why Chinese learners of English are likely to drop the plural suffix *-s* in sentences such as *I have mixed feeling about going home* (compare section 6.5). Since *feeling* represents an abstract concept for them and is not preceded by a classifier, they may tend to leave the noun unmarked.

Hong Kong English

Also for HKE, the use of the plural suffix with StE uncountable nouns is relatively well researched. According to Setter et al. (2010), "[i]n Hong Kong English, the bare form of a noun is normally used for generic reference regardless of whether the noun is a count noun or a mass noun" (45; compare Ziegeler 2015 on SgE above). In StE, in contrast, singular countable nouns are either preceded by a definite or indefinite article, or they occur in the plural form, as described above.

Setter et al. (2010) make an important point by saying that "the semantics of the nouns alone cannot determine countability; countability is more of a grammatical feature than a semantic one in Standard English" (59). The nouns *idea* and *bread* are given as examples, *idea* being semantically abstract but grammatically count and *bread* being semantically concrete but grammatically mass or uncount. With reference to Liu et al. (2006: 136; see above), Setter et al. (2010) stress that Chinese learners of English struggle with English uncountable nouns like *furniture* or *bread* "because they tend to categorise count and mass nouns in terms of the semantics of the nouns" (60). Even more, "in Cantonese, almost everything can be counted through a classifier system" (ibid.). Thus, although a semantic distinction between count and mass nouns exists, it is not realized grammatically. The authors interpret utterances such as *yes and you see . . . there there will be giraffe* (09-MT:03:30[72]) as instances of generic reference. As they point out, "the semantics of [this] noun[ ] has priority in determining the grammar" (ibid.: 61).

According to Wong (2017), "[t]he breakdown of count/mass noun distinctions in HKE can [. . .] be traced back to the syntax of the substrate [Cantonese]" (13). Elements of the Cantonese noun phrase occur in the order demonstrative, numeral, classifier, adjective, noun, and two main functions of classifiers (CL) are to enumerate

[72]The language data Setter et al. (2010: 9) collected stem from students from Hong Kong who studied at the Universities of Reading and Oxford, both UK, at the time of data collection. "MT" stands for map task, a task in which the students were asked to co-operatively guide the interviewer on a map along a route only they saw. The example is from speaker 9's file at three minutes and 30 seconds.

nouns (e.g., *loeng5 zek3 daan2*[73] two CL egg "two eggs") and to individuate them (e.g., *ni1 zek3 daan2* this CL egg "this egg"; cf. Matthews & Yip 1994: 92). Wong (2017: 13) presents two respective types of classifiers. Measure classifiers "denote plurality or uncountable substances" (ibid.: 13) and "can be used to denote both count and mass nouns (e.g., *di1 jan4* CL person 'the/some person' versus *di1 seoi2* CL water 'the/some water')" (ibid.). Type classifiers, in contrast, "reflect intrinsic features of the nouns with which they belong" and "can be used to denote count nouns only (e.g., *ni1 go3 jan4* this CL person 'this person')" (ibid.). What does that mean for the "absence of count/mass noun distinctions in HKE" (ibid.) though? On the one hand, HKE speakers are likely to struggle with the count/mass distinction in English because classifiers which indicate the respective count or mass reference are lacking in English. I.e., there is is no classifier in English that tells them whether the noun is count or mass. On the other hand, the fact that one and the same measure classifier can be used for both count and mass nouns in Cantonese shows that the count/mass noun distinction is not as clear cut in Cantonese as it is in English, after all. Resulting from that, HKE speakers attach the plural suffix to StE mass nouns (e.g., *When people hear interior design some people think oh what's it got to do with furnitures right*, ICE-HK:S2A-058#75:1:A). HKE speakers (like most L2 learners of English) obviously need to study English mass nouns by heart.

Indian English

According to Sedlatschek (2009), "variability along the count-noncount and singular-plural divides has been claimed to be a peculiarity of the IndE noun phrase" (227). Sailaja (2009: 64) mentions variability in count versus non-count uses as a typical feature of IndE as well. Besides uncountable nouns that are treated as countable nouns, unmarked plurals (e.g., *aircraft*) take the plural suffix, and plural nouns occur in the singular (e.g., *a trouser*). Sedlatschek (2009: 228) points out that this is little surprising given the fact that even the "prestige varieties" (ibid.) BrE and AmE vary in this respect (e.g., Schneider 2007: 85). Sahgal & Agnihotri (1985: 126–127) have described countability and number as relatively acceptable features, which is why "many of the forms [...] might have turned into stable usage patterns by the turn of the millennium" (Sedlatschek 2009: 228).

On the basis of a list of 78 nouns discussed in previous literature on IndE (Nihalani et al. 1979; 2004; Yadurajan 2001), Sedlatschek (2009: 229) investigates the *Primary*

[73] For details on the Yale system that is used to represent Chinese tone contours in Roman script, see Matthews & Yip (1994: 7).

Corpus and the *Kolhapur Corpus*[74] for the behavior of the selected nouns. The majority of the nouns have dual-class membership, meaning they occur both with count and non-count uses in StE. While 23 of the 78 selected nouns show the non-standard behavior investigated at least once, Sedlatschek (2009: 229) admits the following:

> [E]ven in the Kolhapur Corpus the token frequencies of the 'different' uses are always lower than those of the standard uses [which is why] one is left with little interpretative leeway apart from stating that contemporary IndE largely follows the codified conventions.

However, a more in-depth look at the *Kolhapur Corpus* and the *Primary Corpus* data reveals that IndE press texts from around 2000 show slightly stronger non-standard noun behavior than press texts from 1978. We certainly need more recent corpus data to determine whether those features have indeed developed towards stable patterns of use.

Sedlatschek's (2009) qualitative insights reveal that it makes sense to distinguish between nouns that cannot take the indefinite article by English standards and nouns that can. With reference to Quirk et al.'s (1985) definition of uncountable nouns as "denoting an undifferentiated mass or a continuum" (ibid.: 246) that "cannot be pluralized and are normally not used with the indefinite article" (Sedlatschek 2009: 231), Sedlatschek (2009: 232) points out the following:

> According to Quirk et al. (1985), the use of the indefinite article in the given contexts must be considered exceptional but not impossible by standard English norms, which allow for the use of *a/an* with noncount nouns when "the noun refers to a quality or other abstraction which is attributed to a person" (287) or when "the noun is premodified and/or postmodified" (287).

In a sentence such as *A huge anger filled me* (IndE KOL K 46), the use of the indefinite article with noncount *anger* "falls within what seems a general (albeit

[74]The *Primary Corpus* comprises 180,000 words of spoken and written English, consisting of press texts, published broadcast material, and student essays (compare section 3.4.2). It was compiled by Sedlatschek in 2000. The one-million-word *Kolhapur Corpus* was designed and compiled by Shastri in 1978 with the intention to match the 1961 LOB (BrE) and *Brown* (AmE; cf. Francis & Kučera 1964) corpora (Sedlatschek 2009: 35).

quite rare) possibility in standard English" (ibid.: 232). *Informations* in *All informations gathered from the agent's report incorporating the surveyor's findings will be important when liability is under consideration*, in contrast, is described as "unusual in standard English" (ibid.). Sedlatschek's (2009) investigation of uncountable nouns in selected collocations by means of Google-based search queries for various top-level country domains (.in, .sg, .uk, .za, .au, .us) is worth mentioning in that context. Examples of the collocations considered are *gave information/gave informations*, *enabling legislation/enabling legislations* and *medical equipment/equipments*. The study shows that IndE does not "differ fundamentally from international usage conventions" (ibid.: 232). Additionally, while Sedlatschek (2009: 232) replicates uses of uncountable nouns as count nouns observed in the *Kolhapur Corpus* and in the *Primary Corpus* by means of his Google searches, the StE alternatives always prevail clearly. For the categorization of countable and uncountable nouns, Sedlatschek (2009) relies on the *Collins Cobuild English Dictionary for Advanced Learners* (Sinclair 2001), for insights into standard uses in earlier decades and centuries he consults the *Oxford English Dictionary* (OED; 1999). Among other things, his analyses reveal that the plural form *equipments* is comparatively frequent in IndE and SgE[75] and that the OED provides evidence of its count use up until 1873 (ibid.: 235). Sedlatschek (2009) concludes that "rather than being a more recent independent Asian innovation proper" (ibid.: 235), it could have "gone out of general use in the major varieties of English but has stayed on more firmly in the Asian region" (ibid.).

7.2.2 Corpus findings

The pluralization of uncountable nouns in the contact varieties of interest was investigated in GloWbE. As Sedlatschek (2009: 229) has shown for IndE, corpora of considerable size are needed to properly investigate the use of uncountable nouns as countable nouns. Due to the strict focus on the pluralization of uncountable nouns (leaving aside zero plurals, plural-only nouns, and collectives), it was possible to account for differences in use among the contact varieties of interest and in comparison with BrE and AmE (see below why AmE was included as a second control variety).

To identify potential candidates for the analyses, a list of the nouns considered in earlier studies on the phenomenon of interest was compiled in a first step. In a second step, further noun candidates were identified by means of the variety-specific

[75] This reasoning implies that the country domain is equivalent to the variety spoken in the country the domain stands for. For cautious words on respective assumptions see section 4.2.

noun lemma frequency lists retrieved from the full-text offline version of GloWbE. The *Oxford Advanced Learner's Dictionary* (Hornby 2005) was consulted to make sure that all nouns identified constitute uncountable nouns in StE.

Table 7.5 provides information on the regularization rates ("reg. rate") in the three target corpora (SG, HK, IN) and in the two control corpora (GB, US) of interest. Only nouns that are regularized at least five times in any of the corpora considered were accounted for. The absolute lemma token frequencies ("lem. freq.") were added because the relative lemma token frequencies that account for differences in corpus size are too small to meaningfully list them in the table. The absolute numbers of uncountable nouns that take the regular plural suffix and of uncountable nouns that occur in the singular are provided in table D.1 in appendix D. The regularization rate for each lemma was calculated by dividing the number of regularized forms by the sum of singular forms and regularized plural forms. The list of potential uncountable candidates for pluralization reveals that many of the lemmata identified are unlikely to occur in spontaneous spoken language (e.g., *machinery*, *immigration*, and *pollution*). This is problematic given the assumption that regularization is more likely in unmonitored than in monitored speech. At the same time, a corpus study based on GloWbE cannot draw conclusions for spontaneous spoken language use, which justifies keeping those lemmata in the sample. In fact, it makes sense to compare nouns that hardly occur in spontaneous spoken language with nouns that are more typical for spoken language (e.g., *information*, *homework*, *furniture*). The latter are probably more likely to be pluralized. Both the uncountable nouns and the regularized countable nouns were searched for by means of the part-of-speech tag ".[n*]" attached to the respective word form of the noun (e.g., "equipment.[n*]," "equipments.[n*]").

As table 7.5 shows, the observed regularization rates are very low. Despite the overall low number of uncountable nouns that take the regular plural suffix, it is worth pointing out that also the control corpora GloWbE US and GloWbE GB contain individual instances of regularization. In many cases the regularization rates in GloWbE US and GloWbE GB are in fact comparable to those in GloWbE SG, GloWbE HK, and GloWbE IN. The general impression gained is that when a uncountable noun occurs with the plural suffix, also verb-noun concord is adjusted, meaning that a present tense verb form does not take the third person singular *-s*. Thus, the uncountable noun is used as a countable noun in all respects. This is true for both the contact varieties and the control varieties.

Table 7.5: Uncountable nouns used as countable nouns in GloWbE (five occurrences or more in any corpus, in alphabetical order, reg. rate in %)

lemma	GloWbE SG		GloWbE HK		GloWbE IN		GloWbE GB	GloWbE US
	reg. rate	lem. freq.	reg. rate	lem. freq.	reg. rate	lem. freq.	reg. rate	reg. rate
advertising	0.04	2,120	0.19	2,481	0.02	4,261	0.03	0.01
advice	2.28	4,446	2.57	4,301	2.12	7,206	0.30	0.42
anger	0.88	992	0.51	752	0.18	3,204	0.20	0.20
assistance	0.06	1,540	0.33	2,243	0.16	3,451	0.13	0.05
blood	0.31	4,653	0.17	3,829	0.19	9,243	0.59	0.21
bread	7.19	2,623	6.16	1,482	9.26	2,050	4.16	4.85
cash	0.02	3,890	0.05	3,510	0.07	7,029	0.02	0.02
consciousness	0.58	642	1.15	1,568	0.19	7,213	0.65	0.54
coordination	0.00	447	0.32	667	0.00	1,381	0.05	0.21
corruption	1.28	961	1.20	1,175	0.88	11,474	0.89	2.23
coverage	0.40	1,403	0.26	2,216	0.55	3,991	0.17	0.91
data	0.02	8,803	0.04	12,900	0.09	32,492	0.01	0.01
education	0.17	8,263	0.22	12,230	0.16	24,980	0.23	0.57
equipment	5.07	2,616	5.45	4,529	8.93	5,697	1.00	0.51
fun	0.05	4,082	0.52	2,538	0.08	4,869	0.07	0.04
furniture	2.47	965	0.74	1,789	0.52	1,811	0.24	0.26
happiness	0.00	2,088	0.00	1,595	0.02	5,465	0.07	0.06
health	0.01	10,151	0.00	10,644	0.03	21,862	0.01	0.00
homework	0.71	542	1.96	628	0.00	1,099	0.48	0.05
immigration	1.25	1,437	0.24	2,746	0.36	1,835	0.19	0.20

(Continued)

lemma	GloWbE SG		GloWbE HK		GloWbE IN		GloWbE GB	GloWbE US
	reg. rate	lem. freq.	reg. rate	lem. freq.	reg. rate	lem. freq.	reg. rate	reg. rate
importance	0.04	2,385	0.00	2,998	0.04	8,946	0.02	0.02
inflation	0.29	1,308	0.28	1,360	0.14	4,609	0.13	0.12
information	0.14	17,801	0.17	24,507	0.23	44,730	0.10	0.08
jewellery	1.18	315	0.00	875	0.81	1,739	0.11	0.32
jewelry	1.33	635	1.15	1,442	3.26	1,571	3.23	0.50
knowledge	0.08	5,726	0.36	7,581	0.08	21,552	0.11	0.11
learning	0.77	3,713	0.38	4,277	2.05	6,531	0.63	0.63
legislation	3.76	826	3.82	1,640	7.90	2,562	0.63	0.44
luck	0.21	2,678	0.13	1,449	0.15	3,643	0.07	0.05
machinery	1.94	386	1.87	1,007	2.23	1,662	0.70	1.03
marketing	0.00	6,825	0.00	6,460	0.00	12,204	0.01	0.02
milk	0.09	3,103	0.14	2,005	0.20	4,782	0.52	0.62
music	0.12	9,788	0.03	10,635	0.03	18,404	0.09	0.07
planning	0.15	1,946	0.00	2,785	0.14	5,161	0.02	0.02
pollution	0.59	634	0.34	2,177	1.17	2,436	0.35	0.62
privacy	0.18	1,592	0.00	2,070	0.03	2,734	0.03	0.10
recognition	0.79	1,322	0.76	2,127	1.29	3,276	0.37	0.54
rice	0.14	4,844	0.28	2,333	0.13	4,216	0.20	0.22
software	0.49	4,041	0.82	4,791	1.37	19,927	0.23	0.32
steel	0.72	915	0.22	2,127	1.04	3,477	0.99	0.93
storage	0.40	1,914	0.41	2,266	1.40	5,109	0.09	0.12
stuff	5.38	5,155	2.80	2,764	3.86	6,436	0.29	0.43
thinking	0.08	2,446	0.69	2,463	0.10	5,710	0.04	0.05
traffic	0.06	3,278	0.20	3,738	0.23	9,032	0.04	0.03

(Continued)

lemma	GloWbE SG		GloWbE HK		GloWbE IN		GloWbE GB	GloWbE US
	reg. rate	lem. freq.	reg. rate	lem. freq.	reg. rate	lem. freq.	reg. rate	reg. rate
training	1.82	6,197	1.60	7,407	1.22	12,448	0.30	0.95
violence	0.00	1,143	0.00	1,411	0.01	8,248	0.05	0.06
weather	0.42	2,658	0.24	2,803	0.27	4,232	0.66	0.09

Recall that GloWbE US was added as a control corpus besides GloWbE GB based on the assumption that variety-specific usage patterns might emerge regarding that feature. While use of uncountable nouns as countable nouns was observed across standard and contact varieties to similar degrees, variety-specific usage patterns in the sense that certain uncountable nouns are relatively much affected in individual varieties were not found.

Most of the lemmata constitute abstract concepts (e.g. *consciousness*, *education*, *happiness*, *knowledge*, and *recognition*) that are per se uncountable and that have very low regularization rates. Other lemmata such as *advice*, *bread*, *furniture*, *legislation*, and *stuff* occur more often with the regular plural suffix *-s*, resulting in higher regularization rates. In examples 7.10 and 7.11, *breads* is used in the sense of "types of bread". The same is true for *advices* (used in the sense of "pieces of advice", example 7.12) and for *furnitures* (used in the sense of "pieces of furniture", example 7.13).

(7.10) Sprouted grain *breads* are less adulterated and more digestible than whole grain *breads* (GloWbE US)

(7.11) This happens to be one of my favorite *breads* I order, when we eat at restaurants (GloWbe IN)

(7.12) Before the end of the presentations, Christoph left a few golden *advices* (GloWbE SG)

(7.13) The choice is large: old *furnitures*, old art pieces, ... (GloWbE HK)

Figure 7.2 depicts the regularization rates for all lemmata in table 7.5 by their logarithmically transformed relative lemma token frequency.

No clear frequency effect is observable. With the exception of a few outliers in each corpus (*legislation*, *equipment*, *bread*, and *stuff* in GloWbE SG; *bread*, *legislation*, *stuff*, *advice*, and *equipment* in GloWbE HK; *bread*, *legislation*, *equipment*, and *stuff* in GloWbE IN), all lemmata have very low regularization rates irrespective of their frequency of occurrence in the corpus.

Due to the very low regularization rates in the target corpora (SG, HK, IN) and their comparability with those in the control corpora (GB, US), it is difficult to account for substratum transfer or institutionalization as determinants of the regularization of uncountable nouns. Regularization rather seems to occur sporadically,

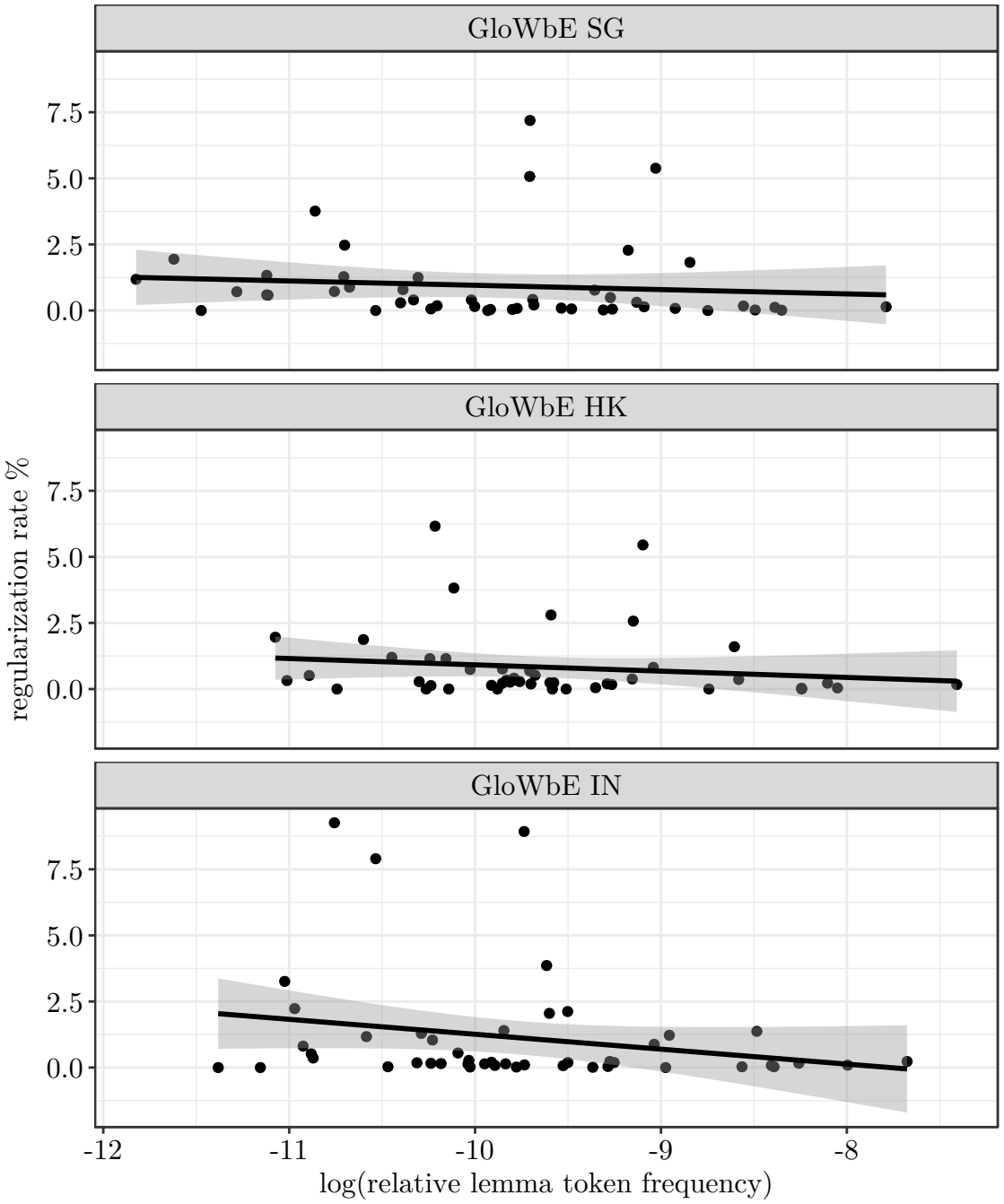

Figure 7.2: Regularization rates in GloWbE by log(relative lemma token frequency), by corpus

lacking any (frequency) pattern. Still, it is likely that different conceptualizations of the countable-uncountable distinction in the various substrate languages at least favor the use of the regular plural suffix with uncountable nouns.

Uncountable nouns occur too seldom with the plural suffix to call this type of regularization an established feature in any of the target varieties. The A-ratings in eWAVE for HKE and IndE as well as the B-rating for CSE are probably due to the salience (rather than the frequency) of the feature.[76] The same was concluded for the regularization of irregular verbs, where it was reasoned that the irregular past tense forms investigated are too entrenched in the mind to be regularized systematically. With uncountable nouns that are treated as countable nouns the matter is a different one though. The use of uncountable nouns as such in StE is partly semantically motivated and partly not, which makes uncountable nouns an error-prone phenomenon for all types of learners of English.

7.3 Concluding remarks

Irregular verbs and uncountable nouns are special cases in their respective word class paradigms because they behave differently from the majority of the word class members. The phenomena were investigated under the assumption that their irregularity is likely to make them prone to regularization (i.e., system-based simplification); particularly in relatively little established varieties such as HKE.

We saw that regularization occurs only sporadically and little patterned; recall the slight frequency effect among the regularized irregular verbs though, with infrequent irregular verbs being more regularized than frequent ones. Those results are certainly influenced by the fact that the analyses concentrated on GloWbE, an internet corpus that is far from providing spontaneous speech material. With the ICE corpora being much too small to investigate the regularization phenomena in a systematic manner, bigger spoken corpora are needed to account for regularization in spontaneous speech.

[76] Rüdiger (2017) refers to the feature as "what might be the shibboleth of non-conventional language use" (104). Her account of "non-count nouns and their use as countable grammatical units" (ibid.) reveals that the feature is rare in English spoken in Korea as well (ibid.: 105–185).

8 Testing the perception of lack of inflectional marking

8.1 Features of interest and hypotheses

This chapter describes a web-based experiment that investigated the perception of lack of verbal past tense and nominal plural marking. The target groups comprised speakers of English from Hong Kong, India, and Singapore, whereas speakers of AmE and of BrE served as the control group. While BrE is the historical input variety of HKE, SgE, and IndE, it can be assumed that participants of the target groups get considerable AmE input from the media in general and from social media in particular. The AmE and BrE participants performed very similarly, which justified treating them as one single control group (see section 8.4). The focus was on the question whether time adverbials preceding the target verbs and quantifiers preceding the target nouns influence perception.

The perception of lack of verbal past tense and nominal plural marking was measured by means of two tasks, namely a self-paced reading task and an acceptability judgment task. In the self-paced reading task, sentences were presented word by word on the screen. Participants read the sentences at their own pace by actively controlling for the one by one display of the words. The previous word disappeared when the next word appeared. The reaction times between pressing the space bar (or tapping the screen when using a touch device) were measured. In the acceptability judgment task, participants saw sentences on the screen (the whole sentence being immediately visible this time) and judged them for their acceptability on a scale from "not acceptable at all" to "fully acceptable." The tasks will be described in detail in the methods section (8.2).

The corpus analyses (chapters 5 and 6) revealed that past tense and plural omission rates in ICE-HK surpass those in ICE-SIN by far, and that ICE-SIN tends to pattern with ICE-IND. Due to the lack of availability of the original ICE recordings, the corpus study on verbal past tense marking focused on instances where past tense marking is not phonologically conditioned, i.e., on vowel-final lemmata

that are followed by a vowel-initial word or pause. This raises the question whether speakers of SgE and HKE differ in their perception of omission of verbal past tense marking as well. Based on the corpus findings, speakers of HKE were expected to deviate particularly strongly from the control group in that they process unmarked verbs and nouns comparatively fast. Still, omission is visible to smaller extents in ICE-SIN and ICE-IND as well and has been described for both SgE and IndE in the literature, which is why also speakers of those varieties were assumed to perceive unmarked verbs and nouns faster than the control group.

The issue is less straightforward with judgments of lack of verbal past tense and nominal plural marking. While speakers of all three contact varieties are more familiar with the features than the control group, this does not necessarily mean that they also evaluate omission of verbal past tense and nominal plural marking more positively. The self-paced reading task was purely performance-based, i.e., it measured the immediate reaction of participants towards lack of inflectional marking. In contrast, the judgments in the acceptability judgment task relied on the metapragmatic assessment of the features and might have been influenced by language ideology. Consequently, it is impossible to disentangle how far the judgments of individuals reflect a common (perhaps society-internal) stance rather than a personal preference. In case the judgments are similar across the target groups, this can be an indicator that omission is tolerated, independent of substratum influence or the degree of institutionalization of English (Alice Blumenthal-Dramé, p.c., 5 December 2016). Overall, the target groups were expected to evaluate omission of verbal past tense and nominal plural marking more positively than the control group because they are more familiar with the features.

Those are the hypotheses underlying the perception experiment that were introduced in section 1.3:

Hypothesis 4 *Speakers of the target varieties (HKE, IndE, SgE) read verbs that lack inflectional past tense marking and nouns that lack inflectional plural marking compared with the mean of the means of all conditions faster than speakers of the control varieties (AmE, BrE).*

Hypothesis 5 *Speakers of the target varieties (HKE, IndE, SgE) evaluate stimuli containing verbs that lack inflectional past tense marking and stimuli containing nouns that lack inflectional plural marking compared with the mean of the means of all conditions more positively than speakers of the control varieties (AmE, BrE).*

As sections 5.3 and 6.3 showed, the low numbers of inflectionally unmarked verbs and nouns (once phonetic environments that promote consonant cluster reduction had been discarded) make it difficult to draw conclusions about the impact of preceding time adverbials and quantifiers on omission. As to the experiment, it was assumed that time adverbials and quantifiers ease the perception of unmarked verbs and nouns following them simply due to the fact that they stress the past time and plural reference, respectively. Again, assumptions concerning the judgment of inflectionally unmarked verbs and nouns preceded by time adverbials and quantifiers were more difficult to make. A time adverbial or quantifier could "remind" participants that the respective unmarked form should be marked (making them more critical of the unmarked form), ease the understanding of the stimulus and be approved accordingly, or lead participants to consider the inflectional affix redundant (resulting in approval as well). As the hypotheses provided above show, the target groups were expected to read stimuli containing unmarked verbs or nouns faster than the control group and to evaluate them more positively. This includes stimuli where the unmarked verb or noun is preceded by a time adverbial or quantifier, respectively. It would not have been feasible to compare all conditions with each other.

8.2 Methods

8.2.1 Participants

Participants were recruited by means of the friend-of-a-friend approach (see section 4.3), via the social media platform Facebook and via email. It sufficed to receive the link to the experiment because the start page provided all necessary background information (purpose of the experiment, target groups, remuneration). Additionally, the link to the experiment was posted in several Facebook groups recommended by participants. Participation was restricted to people above the age of majority. No further restrictions were placed on age, sex, a participant's language background, or the level of education. Participants were told that the project deals with how English is developing around the world and that it investigates how certain language features are perceived by people from different places.

Participants who took part in the experiment could enter a raffle to win the first prize (100 euros) or one of 30 payments of 15 euros by leaving their email address. Leaving one's email address was voluntary but necessary for entering the raffle. The payments were sent via PayPal and converted to the currencies of the winners.

Participants who provided their email addresses automatically received a random six digits code that they could forward to friends. For each friend who was successfully recruited and who provided the code at the end of the experiment, the email address of the participant was automatically put in the raffle one additional time, which increased the participant's chance to win. This was communicated accordingly. The email addresses were stored separately from the experimental data to guarantee anonymity. Participants were told to do the experiment in a quiet place to be able to focus on the tasks.

In case participants paused between task 1 and task 2 or paused for a long period of time within task 1 (20 minutes or more), task 2 was excluded from the analyses. The reason is that the same stimuli were used for both tasks and a pause in between tasks might impact have impacted the judgments of the stimuli. Additionally, whenever it took participants longer than five minutes to work on a stimulus, that same stimulus, the following stimulus, and the second but next stimulus were excluded from the analyses to account for the time needed to get accustomed to the task again.

Table 8.1 depicts the number of participants who took part in the self-paced reading task and in the acceptability judgment task. In order to do the acceptability judgment task, participants had to finish the self-paced reading task first. The acceptability judgment task counts fewer participants because some participants did not proceed to the second task or because the second task could not be considered due to long pauses in the first task (see above).[77]

Table 8.1: Number of participants by speaker group and task

speaker group	self-paced reading (task 1)	acceptability judgment (task 2)
SgE	62	50
HKE	36	27
IndE	52	41
control	43	36
sum	193	154

Table 8.2 summarizes the key background information participants provided. The HKE and control group speakers are on average older than the SgE and IndE speak-

[77] 483 participants dropped out before even finishing the self-paced reading task. This figure comprises all participants who clicked the "start" button to start the experiment but did not continue until the end of the self-paced reading task. It is proof of the high drop-out rates in web-based studies. The responses provided by those participants were not counted.

ers, and across speaker groups (except for the IndE speakers), more female than male respondents participated. Particularly noteworthy is the young mean age at which the HKE speakers started learning English (mean age English); also in comparison with the IndE speakers. A look at the home languages of participants reveals that about one fourth of the SgE and HKE speakers reported speaking exclusively English at home (table 8.5). The IndE speakers nearly exclusively reported to have acquired a language other than English as their first language. The language background information participants from Hong Kong provided is surprising. Possibly, mainly HKE speakers with a certain confidence in their English skills because of having reached as certain proficiency level decided to take part in the experiment. Compared with the SgE and HKE speakers, the IndE speakers and the control group have reached comparatively high levels of education. The majority of the respondents across groups used a laptop to do the experiment and have not attended linguistics classes. Overall, none of the key background variables are strongly over- or underrepresented in any of the speaker groups, which was important for model fitting. In case different predictor variables in a regression model strongly correlate, one speaks of "multicollinearity" (cf. Baayen 2012: 37). When multicollinearity is an issue, it is impossible to disentangle the impact of single predictor variables on the dependent variable, which is problematic. The language background of the participants will be elaborated on in the analyses in section 8.4.[78]

8.2.2 Materials

Tasks and procedure

Participants were first presented with the self-paced reading task and then with the acceptability judgment task. In the self-paced reading task, they read sentences on the screen. The sentences were presented word by word in line with the so-called "subject-paced moving window paradigm" (Just et al. 1982: 230), i.e., the words successively appeared next to each other as they would in a normal text (see figure E.1 in appendix E.3 for the task design). Participants actively controlled for the presentation of each word by pressing the space bar (or tapping the screen, when they used a touch device) once they were ready for the next word. Upon display of the next word, the previous word disappeared again. Periods and commas were

[78] See appendix E.1 for the background questionnaire participants were presented with. Appendix E.2 lists the follow-up questions participants saw upon completion of the acceptability judgment task.

Table 8.2: Key background information on participants by group and task

group & category	self-paced reading task	acceptability judgment task
SgE		
mean age	28.03	27.51
sex	female: 48, male: 13, not answ. 1	female: 39, male: 10 , not answ.: 1
mean age English	2.10	2.10
level of education	< Bachelor's: 23, Bachelor's: 25, ≥ Master's: 13, not answ.: 1	< Bachelor's: 19, Bachelor's: 20, ≥ Master's: 10, not answ.: 1
device	laptop: 39, smartphone: 23	laptop: 33, smartphone: 17
linguistics classes	no: 49, yes: 12, not answ.: 1	no: 39, yes: 10, not answ.: 1
HKE		
mean age	34.63	34.78
sex	female: 27, male: 8, not answ.: 1	female: 21, male: 6
mean age English	2.38	2.62
level of education	< Bachelor's: 9, Bachelor's: 18, ≥ Master's: 8, not answ.: 1	< Bachelor's: 5, Bachelor's: 15, ≥ Master's: 7
device	laptop: 18, smartphone: 18	laptop: 15, smartphone: 12
linguistics classes	no: 28, yes: 7, not answ.: 1	no: 21, yes: 6
IndE		
mean age	28.14	27.70
sex	female: 22, male: 29, not answ.: 1	female: 17, male: 23 , not answ.: 1
mean age English	4.24	4.40
level of education	< Bachelor's: 6, Bachelor's: 16, ≥ Master's: 29, not answ.: 1	< Bachelor's: 4, Bachelor's: 12, ≥ Master's: 24, not answ.: 1
device	laptop: 34, smartphone: 18	laptop: 28, smartphone: 13
linguistics classes	no: 38, yes: 13, not answ.: 1	no: 29, yes: 11, not answ.: 1
control		
mean age	35.91	37.06
sex	female: 24, male: 19	female: 18, male: 18
mean age English	0.87	1.03
level of education	< Bachelor's: 12, Bachelor's: 11, ≥ Master's: 19, not answ.: 1	< Bachelor's: 11, Bachelor's: 10, ≥ Master's: 14, not answ.: 1
device	laptop: 21, smartphone: 22	laptop: 20, smartphone: 16
linguistics classes	no: 33, yes: 9, not answ.: 1	no: 28, yes: 7, not answ.: 1

presented with the words that preceded them. The words appeared successively on a line that visualized the sentence length. Since screen size cannot be controlled for in web-based experiments, the line was crucial for orientation because it made sure that participants knew when to expect a line break. With their subject-paced moving window paradigm, Just et al. (1982: 230) observe word-level effects known from studies measuring the eye fixation of subjects that read normal text: Readers pause comparatively long on lengthy or less frequent words, when a new topic is introduced, and sentence-finally. Alternative designs to the moving window paradigm are the "stationary window paradigm" and the "cumulative window paradigm." In the stationary window paradigm, each word is presented in the same position on the screen (i.e., in the middle) rather than next to the previous word. This design was decided against here because it does not mimic actual reading of English sentences from left to right. In the cumulative window paradigm, words are presented successively next to each other (as in the moving window paradigm), but the previous word does not disappear when the next word appears. This design was not used here because it was considered too easy. The reaction times, i.e., the times between pressing the space bar (or tapping the screen), were measured.

The presentation of the stimuli was preceded by instructions (including a short tutorial that demonstrated participants how to navigate through the task) and a practice session with three practice sentences that did not count towards the task. Participants were instructed to read the sentences at their own pace and were told that upon pressing the space bar (or tapping the screen) the currently visible word would disappear and the next word would appear. They were informed that some of the sentences would be followed by a comprehension question and were instructed to answer the comprehension question as quickly and accurately as possible. Each comprehension question was a yes/no question, and participants provided their answer by clicking the respective button. If the answer was correct, the button turned green, if it was wrong, the button turned red. Thus, participants received immediate feedback on their answer choice. The comprehension questions had to be answered in order to be able to continue and served two purposes: Firstly, they reminded participants to read the stimuli properly in order to be able to answer the comprehension questions correctly. Secondly, participants who did not answer at least 80 percent of the comprehension questions correctly were excluded from the analysis of the task.

In the acceptability judgment task, participants were again presented with sentences on the screen. This time, the whole sentence was visible at once and participants were asked to judge it for its acceptability on a scale from "not acceptable at all" to "fully acceptable" (see figure E.2 in appendix E.3 for the task design). Participants were instructed to imagine that each sentence was uttered by a friend and to judge whether the respective sentence made that friend sound like a native speaker of English. The more the sentence makes the friend sound like a native speaker of English, the more acceptable participants should indicate the sentence to be. "English" was not specified further (e.g., as a "variety of English the participant is used to hearing") because this might have biased the target groups towards evaluating the sentences that contain familiar non-standard features particularly negatively; knowing that they take part in an experiment. The literature recommends not telling participants to judge the degree of grammatical correctness of a sentence, the likelihood of actually hearing such a sentence, or the truth of its contents (cf. Schütze & Sprouse 2013), which is why those aspects were not mentioned in the instructions.

To provide their judgment, participants placed a slider on a seemingly continuous horizontal scale from "not acceptable at all" to "fully acceptable," which they found right below the sentence. The horizontal scale represented 100 points from zero ("not acceptable at all") to 100 ("fully acceptable"), so each position of the slider on the scale represented a numeric value that was used for the analyses. The scale was not visible to participants. In the pilot phase, a swapped scale had been tested for a few distractor stimuli, meaning that the horizontal scale ranged from "fully acceptable" to "not acceptable at all" for those stimuli. Some pilot participants mentioned they did not keep checking the scale for each single sentence, got confused when noticing the swapped scale, and wondered how many swapped scales they had missed. The intention of using a few swapped scales had been to remind participants to concentrate on the task, but the idea was dismissed because of the confusion they caused. As in the self-paced reading task, half of the sentences were followed by a yes/no comprehension question (for an example, see figure E.3 in appendix E.3).

Further contents

The tasks were preceded by a statement of informed consent, information on technical aspects, a number of background questions, and followed by a few follow-up questions. The statement of informed consent comprised information on the contents of the experiment (background questions, self-paced reading task, acceptability judgment task, and follow-up questions) and the remuneration of participants, as well as

on data storage, protection, and deletion.[79] The experiment software was explicitly programmed for the purposes of the experiment (see section 4.3), and the experimental data are stored on https://uberspace.de, a German server host. The data are anonymous because the server host only saved a short and therefore anonymized version of a participant's IP-address (providing information about country and region only) in addition to the participant's operating system and browser. From the experimental data and email addresses, a reference list was created that is stored separately and that makes it possible to delete the experimental data of participants who decide to withdraw from the experiment later on. Participants had to confirm that they accept the statement of informed consent and that they had reached the age of majority before they could continue. They were told not to use the navigation functions of their browser and that back navigation is only possible when there is a back button (e.g., in the task instructions). Furthermore, participants were informed that their current position in the experiment was being saved continuously. In case participants accidentally closed the browser tab or the internet connection was interrupted, they could open the link to the experiment again and were redirected to their last position. All actions throughout the experiment were time logged so respective reloads could be traced back.

The background questions covered the participants' age and sex, the country they currently live in, their total time spent abroad and in English-speaking countries (not including holidays), their language skills, highest completed level of education, and current education status (e.g., undergraduate student or graduate student). Participants were asked about their native language, other languages and/or dialects they speak (provided in the order of proficiency), the age at which they started learning English, their home language(s), and their handedness. Singaporean participants were additionally asked to indicate the ethnic group they belong to (Chinese, Malay, Indian, Other). Answering the background questions was optional, but participants were told that providing the respective information was of importance for the analyses.

A short group task preceded the background questions in which participants had to indicate where they grew up. If they were born in one country or region but raised in another, participants were asked to choose the place they spent most of their childhood in. The answer options were Great Britain, Hong Kong, India, Singapore,

[79] Data storage, protection, and deletion was discussed with members of the ethics department of *Universitätsklinikum Freiburg* and with the *Stabstelle Allg. Rechtsangelegenheiten, Stiftungs- und Vermögensverwaltung/Steuern* of the University of Freiburg.

US and Other, and clicking an option assigned participants to the respective speaker group. Depending on the speaker group they had been assigned to, participants received one of four stimulus lists (for details see subsection **Design**). Participants who clicked "Other" were told that they were not eligible to take part and could not continue the experiment.

At the end of the experiment, participants were presented with a few follow-up questions. They could indicate whether they had an idea what the tasks had been about, and whether it was a topic they were familiar with. Participants were also asked to indicate how confident they felt when they provided their judgments in the acceptability judgment task. Additionally, they could leave further comments.

Design

Each participant saw 84 stimuli (56 target stimuli and 28 distractor stimuli) in the self-paced reading task and 60 stimuli (40 target stimuli and 20 distractor stimuli) in the acceptability judgment task. Table 8.3 provides the critical conditions and the control conditions for set types V (verbs) and N (nouns) with example sets, the critical conditions being those in which the target verb or noun is not inflectionally marked (conditions V2, V4, N2, and N4). The stimuli in each set were identical except for the condition they occurred in.

The same sets were tested in both the self-paced reading task and the acceptability judgment task to find out how participants judged the sets they had been presented with in the self-paced reading task. Participants who received one of the critical conditions for a set in the self-paced reading task received one of the control conditions (V1, V3, N1, N3) for the same set in the acceptability judgment task and vice versa. Since the group task preceding the background questions assigned participants to one out of four predefined stimulus lists, all stimuli in all conditions were equally distributed within the speaker groups. Each list had a unique order of conditions for the self-paced reading task and for the acceptability judgment task that was the same for set types V and N. The order of conditions was obtained by means of a Latin square design. Participants assigned to list 1, for instance, received condition 1 for set 1, condition 2 for set 2, 3 for 3, 4 for 4, 1 for 5, 2 for 6, etc. With 28 stimuli in the self-paced reading task and 20 stimuli in the acceptability judgment task, participants received seven stimuli per condition in the self-paced reading task and five stimuli per condition in the acceptability judgment task. For each set, condition 1 was paired with condition 4 and condition 2 with condition 3 (see table E.1 in appendix E.3 for more details).

Table 8.3: Conditions for set types V and N

conditions	example set
set type V:	
V1: time adverbial, target verb inflectionally marked for past tense	Recently Isabel and Lenny moved away and died in a car crash close to their new home.
V2: time adverbial, target verb not infl. marked for past tense	Recently Isabel and Lenny moved away and die in a car crash close to their new home.
V3: no time adverbial, target verb infl. marked for past tense	Isabel and Lenny moved away and died in a car crash close to their new home.
V4: no time adverbial, target verb not infl. marked for past tense	Isabel and Lenny moved away and die in a car crash close to their new home.
set type N:	
N1: quantifier, target noun inflectionally marked for plural	Bill told us many details about the terrible accident he had been involved in.
N2: quantifier, target noun not infl. marked for plural	Bill told us many detail about the terrible accident he had been involved in.
N3: no quantifier, target noun infl. marked for plural	Bill told us details about the terrible accident he had been involved in.
N4: no quantifier, target noun not infl. marked for plural	Bill told us detail about the terrible accident he had been involved in.

As mentioned in the task descriptions, 50 percent of the sets in set types V and N were followed by a comprehension question that was the same for all four conditions. Thus, each participant received the same comprehension questions per task, which made performances in answering the questions directly comparable. The sets that had not received a comprehension question in the self-paced reading task received one in the acceptability judgment task and vice versa.

The self-paced reading task comprised 28 distractor stimuli and the acceptability judgment task 20 distractor stimuli (set type D). Those stimuli contained non-standard grammatical features other than lack of inflectional past tense or plural marking, semantically awkward lexical material, or none of the two. The non-standard grammatical features chosen were object pronoun drop, subject pronoun drop, and conjunction doubling (for details on the selection process see subsection **Stimuli**). Table 8.4 provides an overview of the different types of distractor stimuli,

their number of occurrences in the two tasks, and an example stimulus each. The distractor stimuli in the self-paced reading task differed from those in the acceptability judgment task, and all participants saw the same distractor stimuli. Half of the distractors (50 percent per distractor feature) were followed by a comprehension question.

Table 8.4: Distractor stimuli (set type D)

condition†: feature	SPR	AJT	example (not semantically awkward)
D5/D6: object pronoun drop	6	4	The two will only buy later this week, so they cannot give feedback yet.
D7/D8: subject pronoun drop	6	4	Have no clue which gym to pick, but I have to do more for my fitness.
D9/D10: conjunction doubling (correlative conj.)	6	4	Although Catherine had a tight schedule, but she took time off.
D11/D12: no feature	10	8	I have the impression that Bob has great potential to excel in his business.
sum	28	20	

†The second condition mentioned per feature (D6, D8, D10, D12) contained additional semantically awkward lexical material in order to make participants pay attention to what they are reading. The numbering starts with 5 for reasons of internal coding.

Each participant received a stimulus list with a unique pseudorandomized order: Stimuli followed by a comprehension question occurred within one to three stimuli. Stimuli of the same set type (V, N, D) occurred within a distance of at least two. In order to keep the critical conditions (V2, V4, N2, N4) as far apart as possible, they did not directly follow each other across set types; even if a distractor occurred in between. Distractors of one feature type did not occur in sequence. The same time adverbials and quantifiers occurred more than once, which was why the pseudorandomization made sure that the same time adverbial or quantifier did not occur within a distance of less than ten.

Both tasks were preceded by three practice stimuli (set type P) each that contained neither grammatically non-standard nor semantically awkward material. Per task, two of the practice stimuli were followed by a comprehension question.

Stimuli

To the author's knowledge, no experimental studies exist that investigate the perception of inflectional past tense and noun plural marking in the speaker groups of interest. The target verbs and target nouns analyzed in the experiment were adopted from the corpus studies in chapters 5 and 6, with the exception of a number of nouns with high plurality rates (see below). Utterances from ICE in which the respective target verbs and nouns occur helped creating the stimuli, but the utterances had to be adjusted to a considerable extent to fit the average sentence length (14.22 words) and to avoid lexical priming.

The 28 target verbs in set type V were *play*, *agree*, *allow*, *stay*, *show*, *follow*, *continue*, *apply*, *enjoy*, *carry*, *die*, *identify*, *worry*, *view*, *issue*, *argue*, *destroy*, *deny*, *employ*, *supply*, *reply*, *rely*, *pray*, *vary*, *cry*, *tie*, *satisfy*, and *bury*. Only verbs ending in a vowel were chosen in order to avoid perception biases among the SgE and HKE participants who might be familiar with inflectionally unmarked verbs ending in a consonant.[80] To be left with enough target verbs, verbs ending in *-y* that form their past tense with *-ied* (e.g., *carry*) were taken into account. The impact of the change in spelling (which can play a role in reading) was considered in the analyses. Furthermore, ten of the target verbs are homographs (*play*, *stay*, *show*, *worry*, *view*, *issue*, *supply*, *reply*, *cry*, *tie*); an aspect which was taken into account as well.

To be able to compare the perception of inflectionally unmarked verbs across speaker groups, it was necessary to use a common frequency basis. It was decided to focus on the lemma frequencies in COCA because no corpus of comparable size exists for BrE, the historical input variety, and because AmE has become increasingly accessible to people around the world due to its strong presence in the (social) media. Lemma (rather than word form) frequencies were chosen because a) word form frequencies differ across the inflectionally marked and unmarked target forms and therefore intermingle with the features of interest, and b) lemma frequencies have been used in the corpus analyses for the reason mentioned in a). Another factor that needed to be considered is word length. The number of letters per word was used as a measure for word length and was adopted from WebCelex (Max Planck Institute for Psycholinguistics 2001).

[80] An analysis of differences in the perception of inflectionally unmarked verbs that end in a vowel and inflectionally unmarked verbs that end in a consonant would go beyond the scope of this experiment but would be worth conducting. It could provide valuable insights into the degree of familiarity of the different speaker groups with lack of the *-ed* suffix in different phonetic environments. Incorporating the various morphological processes involved in the past tense marking of irregular verbs would have gone beyond the scope of the experiment as well.

In the acceptability judgment task, only 20 sets of set type V were needed. Of the eight sets not used for the acceptability judgment task, six contained a homograph as the target verb. The other two sets were randomly selected (compare table E.2 in appendix E for an overview of the stimuli used in task 1 and task 2).

Each target verb was preceded by another regular verb marked for past tense. This was necessary because in sentences with no preceding time adverbial the past time reference needed to be provided. The potential priming effect arising from the preceding marked verb was the same for all four conditions though. Each sentence consisted of two clauses joined by the conjunctions *and* or *but*. In order to avoid sentence wrap-up effects, neither the target verbs nor the lexical material immediately following them occurred at the end of the sentence or close to the end. It has been claimed that sentence or clause wrap-up effects occur because of integrative processing, which means that sentence- or clause-final processing is "involved in relating sentences or clauses and updating a discourse model" (Just & Carpenter 1980, in Warren et al. 2009: 132) and therefore results in longer reaction times.[81]

The sounds following the target verb with past time reference were also controlled for. Gut (2009a: 147–148) notes that word-final plosive deletion is particularly likely in SgE when the following word begins with a nasal or lateral and particularly unlikely before a pause or a non-sibilant fricative (nasal, lateral > plosive > sibilant > glide > vowel > non-sibilant fricative > pause). With a few exceptions, the critical verbs and the verbs preceding them in the first clause were followed by a word beginning with a vowel or a semivowel.

In conditions V1 and V2, a time adverbial occurred sentence-initially. As elaborated on in section 5.3, time adverbials present different time-related meanings (cf. Biber et al. 2007: 777), such as a position in time (e.g., *yesterday*), or the duration of an event (e.g., *for years*). While several kinds of time adverbials co-occured with the sampled verbs in the corpus data, the experiment focused on time adverbials that describe a position in time exclusively. This excluded potential variation in the impact the time adverbial has on the processing of verbs that are not inflectionally marked for past tense. The adverbials used were *last X*[82], *yesterday*, *X ago*, *the other day*, *back then*, *recently*, and *in X*, each occurring in more than one set.

The 28 target nouns in set type N were *year*, *thing*, *problem*, *friend*, *day*, *hour*, *student*, *eye*, *parent*, *term*, *detail*, *shoe*, *school*, *point*, *way*, *question*, *teacher*, *reason*,

[81] Compare Warren et al. (2009) for a discussion of the integrative processing hypothesis that investigates a potential interaction between sentence complexity and wrap-up.

[82] *X* is a placeholder here.

girl, *boy*, *ear*, *toe*, *finger*, *hand*, *member*, *idea*, *game*, and *area*. As with the target verbs, attention was paid to the final sound of the noun (i.e., the sound that precedes the regular plural suffix). On the basis of the findings by L. Lim (2004: 33) and Gut (2005: 20–22) described in section 6.1, nouns that end in a plosive which is not part of a consonant cluster were not considered, whereas nouns that end in three-consonant clusters with one plosive sound were part of the sample. In such three-consonant clusters it is the plosive that is most likely to be deleted, leaving the plural *-s* unattached. Nouns that end in a sibilant and therefore attach another syllable to mark plural were not included because they are not comparable to nouns that attach no syllable.

Four of the sampled nouns are homographs (*term*, *point*, *question*, and *reason*), and the sampled nouns differ in word length as well as in their lemma frequencies. Nouns that form their plural in an irregular way could not be considered here but would be worth investigating. Only 20 sets of set type N were needed for the acceptability judgment task. Of the eight sets not used, four contained a homograph as the target noun. The other four sets were randomly chosen (compare table E.2 in appendix E).

The critical nouns did not occur at sentence boundaries and were preceded by a quantifier in conditions N1 and N2. Only quantifiers were used as indicators of plural reference because in the corpus study the share of nouns not inflectionally marked for plural turned out to be highest after a quantifier (compare section 6.3, table 6.5). Additionally, the focus on one determiner type reduced potential variation in the impact of preceding determiners on the processing of the inflectionally unmarked nouns. The quantifiers used were *several*, *a lot of*, *a few*, *all*, *many*, *various*, and *both*, and each quantifier occurred in more than one set. In all sets, the target noun were in object position because many of the target nouns used are unlikely to function as agents.

As mentioned above, some of the distractor stimuli comprised non-standard grammatical features other than the features of interest, some contained semantically awkward lexical material, and others contained neither of the two. The distractor stimuli with other non-standard features were supposed to distract from the features of interest, whereas the semantically awkward stimuli and those with no particular features made sure that not too much weight was placed on non-standard features. The following non-standard features were chosen on the basis of eWAVE: object pronoun drop (feature 42), subject pronoun drop: referential pronouns (feature 43), and conjunction doubling: correlative conjunctions (feature 215). Those features are

no verb or noun features and should be familiar to the target groups. Object pronoun drop is categorized as "pervasive or obligatory" (rating A) in all three contact varieties (HKE, IndE, and CSE); the same is the true for subject pronoun drop. Conjunction doubling is rated as "pervasive or obligatory" in HKE and as "neither pervasive nor extremely rare" (rating B) in CSE and IndE.

The comprehension questions were content-based yes/no questions and the same comprehension question was used across conditions. The example set for set type V presented in table 8.3, *(Recently) Isabel and Lenny moved away and die(d) in a car crash close to their new home*, for instance, was followed by the comprehension question *Did Isabel die in a car crash?* (Yes). The example set for set type N (*Bill told us (many) detail(s) about the terrible accident he had been involved in*) was accompanied by the question *Had Bill been involved in an accident?* (Yes).

8.3 Pilot study

The pilot phase consisted of three steps. In a first step, the stimuli were proofread by a native speaker of BrE and a speaker of English from Singapore, and the experiment software was tested for user-friendliness as well as data security. The software testers clearly preferred the presentation of the words in the self-paced reading task side-by-side, and not in-place. The acceptability judgment task (which had originally consisted of 84 stimuli as the self-paced reading task) took considerably longer than the self-paced reading task, which is why it was reduced to 60 stimuli in order to meet the time frame of approximately 30 minutes that participation was supposed to take. In a second step, six participants (two from Singapore, four from Hong Kong) were tested on previous versions of the stimuli with the intention to test the changes made. In a third and last step, the finalized stimuli were tested on twenty participants (ten from Singapore, one from Hong Kong, four from India, five from the control group). The data of the twenty participants who participated in the pilot study counted towards the final study because no changes were made to the experiment design or the stimuli after those participants had taken part.

Figure 8.1 plots the 95 percent confidence intervals for the reaction times (left; self-paced reading task) and judgments (right; acceptability judgment task) by condition. Each confidence interval is depicted symmetrically around the respective sample mean (cf. Baayen 2012: 75) and indicates the region which the true population parameter would fall into in 95 percent of the cases were the same population

tested numerous times. Conditions V2, V4, N2, and N4 constitute the "unmarked" conditions, i.e., here the target verb or noun is not marked for past tense and plural respectively. In V1 and V2, a time adverbial precedes the marked (V1) and unmarked (V2) verb, and in N1 and N2 a quantifier precedes the marked (N1) and unmarked (N2) noun. In V3 and V4, no time adverbial precedes the marked (V3) and unmarked (V4) verb, whereas in N3 and N4 no quantifier precedes the marked (N3) and unmarked (N4) noun. As elaborated on in section 8.4 below, in the self-paced reading task, the reaction times towards the target verb or noun and the three words directly following it (called "critical region") rather than towards the target form exclusively were of interest. Figure 8.1 does not distinguish between groups (SG, HK, IN, control).

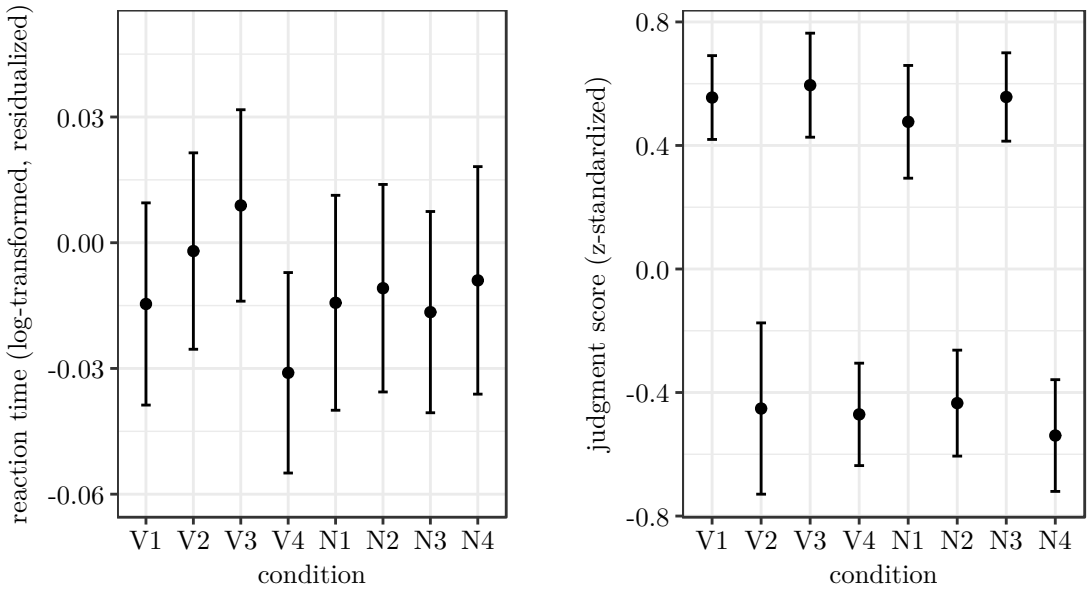

Figure 8.1: 95 percent confidence intervals for reaction times and judgment scores by condition (pilot study)

The reaction times in the noun conditions follow the expected patterns. I.e., participants read the critical regions in conditions N2 and N4 comparatively slowly. The picture is less clear for the verb conditions. The critical regions in condition V2 (unmarked, time adverbial) were read slower than in V1 (marked, time adverbial), but those in condition V4 (unmarked, no time adverbial) were read fast and those in condition V3 (unmarked, time adverbial) comparatively slowly. The difference between conditions V3 and V4 disappeared with increasing amounts of data. Concerning the judgment scores, the "unmarked" verb and noun conditions (V2, V4,

N2, N4) were evaluated much more negatively than their marked counterparts, irrespective of the presence or absence of a time adverbial or quantifier. The more positively a stimulus was evaluated, the higher its judgment score is.

8.4 Results

8.4.1 Self-paced reading task

The assumption underlying the self-paced reading task was that speakers of the target varieties read verbs that lack inflectional past tense marking and nouns that lack inflectional plural marking faster than the control group because of their familiarity with the features (see hypothesis 4, sections 1.3 and 8.1). A number of rules were applied to exclude reaction times from the analyses that are likely to have resulted from disturbances, lack of attention, etc. Firstly, reloaded stimuli were removed from the analyses. Secondly, participants who answered 80 percent or less of the comprehension questions correctly were not considered (cf. Jiang 2012: 70). Thirdly, all stimuli whose comprehension question had not been answered correctly were removed. Reaction times higher than 2000 ms and lower than 50 ms were not considered either (ibid.).

Initially, only the reaction times towards the target verbs and nouns had been of interest. However, it was observed that in the critical conditions (V2, V4, N2, N4) the target forms themselves were not processed considerably more slowly than their marked counterparts. An investigation of the raw reaction times towards the target forms and the three words directly following them showed that reaction times tended to be delayed. I.e., longer reaction times were observed for the words following the inflectionally unmarked target forms rather than for the target forms themselves. This is why it was decided to add the reaction times towards the words following the target forms to the analyses. To account for the differences in word length and lemma frequency of those words, residualized reaction times were used: For each participant, the impact of the number of letters of a word (`Letters`), the position of the stimulus in the task (`ResponsePosition`), and the position of the word in the stimulus (`WordPosition`) on the logarithmically transformed reaction times (`logRT`) was estimated by means of a linear mixed-effects model for all words the participant was presented with in the self-paced reading task (excluding the practice items). The residuals gained were then used as residualized reaction times for the analyses. The model was adopted from Jaeger's blogpost (cf. Jaeger 2008)

and adjusted. The residuals were extracted from the model by means of the function `residuals(model.r)` and added as a variable `logRTresidual` to the data frame `df.spr`. This is the linear mixed-effects model (`model.r`) used:

```
lmer(logRT ~ Letters + log(ResponsePosition) + log(WordPosition)
             + (1|Participant), df.spr, na.action=na.exclude)
```

Preparing the model

Let us elaborate on the independent variables that were used for model fitting. Because of the many background variables that potentially impact on the reaction times measured, a criterion-based approach with backward selection was applied. When backward selection is used, a maximal model that comprises all (theoretically meaningful) independent variables and their interactions serves as the starting point (Gries 2013: 259).[83] Successively, predictors without significant contribution to the model are discarded. Elimination starts at the highest level of interactivity and proceeds to the main effects. The Akaike Information Criterion (AIC) was used to determine the explanatory value of removing or adding predictors (criterion-based approach; Gries 2013). The AIC "relates the quality of a model to the number of predictors it contains" (ibid. 261). Given two models with equal explanatory value, the AIC of the model with fewer predictors is smaller than that of the model with more predictors. As long as the AIC does not increase, it is therefore valid to remove or add predictors.

The dependent variable was the residualized and logarithmically transformed reaction time (`logRTresidual`), as described above. The predictor variables are presented on the following pages.[84] For the categorical variables, the reference level is additionally provided. The reference level is the level that serves as a means of comparison for the other factor levels the respective predictor variable has when dummy coding is used. Dummy coding is the default coding scheme in R. With dummy coding, the mean value of each factor level is compared with the mean value of the reference level of that factor. Let us apply this to the variable `Group`, before we continue with the other predictor variables.

[83] The analyses conducted here were strictly limited to theoretically meaningful independent variables and to theoretically meaningful interactions.

[84] The way of presenting the predictor variables was inspired by Horch (2017: 212–215). The same is the case for the grouping of the first languages participants provided (ibid.: 211), and for the way of presenting the model output (ibid.: 218).

Group The group a participant belongs to was determined on the basis of the response to the group task preceding the background questions (see subsection **Further contents** above). The group task asked where participants had spent most of their childhood. The control group (Control) served as the reference level because this group was expected to behave most conservatively, i.e., that it takes particularly long to read the unmarked critical regions. The mean reaction times of all other groups (i.e., the factor levels SG, HK, and IN) were compared with the mean reaction time of the control group. The participants are referred to as control group speakers, speakers of SgE, speakers of HKE, and speakers of IndE in the following.

Condition `Condition` marks the type of stimulus participants were presented with. The focus was on set types V and N and the four conditions therein (see table 8.3). In the analysis of the self-paced reading task, the distractors were not considered because some of them contained additional semantically awkward material (see the footnote in table 8.4), whose impact on the reading times for the critical regions would have been impossible to determine. Recall, however, that each participant's reaction times towards the words in the distractor stimuli were considered for calculating the residualized reaction times. This predictor variable was sum coded to compare the mean residualized and logarithmically transformed reaction times for a condition with the mean of the means of all other conditions. In contrast with dummy coding, the estimate for sum coded factors indicates the difference between the grand mean (here, the grand mean reaction time) and the respective factor level (here, the reaction time in a particular condition; cf. Granena et al. 2016: 110).[85] Condition N1, which was read fastest across groups, served as the reference level, but in contrast with dummy coding it is not the factor level the other factor levels are compared with. The interaction of `Condition` with `Group` was used to gain insights into the differences in performance by group.

Frequency The logarithmically transformed lemma token frequency, a continuous variable, was derived for each word from COCA for the reasons mentioned in subsection **Stimuli**. As pointed out there, it was necessary to use a common frequency basis for the analyses in order to be able to compare the perception of inflectionally unmarked verbs across speaker groups. In the analysis of the self-paced reading task, the frequency of the critical word plus the three words following the critical word

[85] Values of 1 and -1 were used for sum coding. Those are the default values assigned to the factor levels in R when dummy coding is changed to sum coding by means of the following command: "contrasts() = contr.sum."

was taken into account. Not only the reaction times towards the critical words but also those for the three words following them were of interest.

SpellingChange As pointed out in subsection **Stimuli**, some of the target verbs end in *-y* and form their past tense with *-ied* (*apply*, *carry*, *identify*, *worry*, *deny*, *supply*, *reply*, *rely*, *vary*, *cry*, *satisfy*, and *bury*). A spelling change might result in longer reading times because the lack of marking is particularly obvious when the marked verb requires a change in spelling. Spelling change (Change) served as the reference level.

Equivalent Additionally, some of the target verbs and nouns have noun or verb equivalents that do not differ in form from the target forms. Reading times might be longer for both marked and unmarked forms that have a noun or verb equivalent. Having an equivalent (Equivalent) was the reference level.

Age Participant age (centered) was included as one of the continuous speaker background variables. Older participants were expected to take longer to read the stimuli and the critical regions therein than younger participants. By means of centering, the overall mean value of `Age` was subtracted from each individual age value in the data set (cf. Gries 2013: 130). Centering is useful when predictors lack a meaningful zero point (as is the case with age) and can circumvent multicollinearity (i.e., high correlation) with other predictors in the model, although the latter point is debated (e.g., Dalal & Zickar 2012). High correlation between predictor variables needs to be avoided because regression models assume independent predictors.

FirstLanguage English served as the reference level because it was assumed that L1 speakers of English behave most conservatively in that they pay particular attention to the unmarked verbs and nouns. Surprisingly many of the SgE and HKE speakers (70.97 percent and 36.11 percent, respectively) indicated that English is the language they learned first (see table 8.5). The question asking for the first language of participants was an open question, and the answers provided were grouped into the categories English, Chinese, Other, English & Chinese, English & Other. The reason to put all languages aside from English and Chinese into the category "Other" is that respondents with a Chinese language background (whatever dialect) were expected to be particularly familiar with lack of inflectional marking.

Sex It was assumed that females would be faster in reading the critical regions than males because females have been repeatedly described to promote language change

(e.g., Labov 2001; Eckert & McConnell-Ginet 2013). Females (Female) served as the reference level.

Education The background question asking for the highest completed level of education was designed as an open question to take account of the different educational paths in the countries the participants grew up in. On the basis of the answers provided, it was decided to focus on a distinction between participants who received a master's degree or higher, those who hold a bachelor's degree, and those who do not hold a bachelor's degree (yet). Participants with a master's degree or higher (Master's degree or higher) served as the reference level based on the assumption that their linguistic behavior would be most conservative.

LinguisticsClasses The background questionnaire additionally asked participants whether they had attended linguistics classes in order to account for background knowledge in linguistics. Having taken linguistics classes (Yes) served as the reference level because it is likely that participants with respective background knowledge pay comparatively much attention to unmarked verbs and nouns. The start page of the experiment explicitly asked for participants who have not taken classes in linguistics, but it turned out that the friend-of-a-friend approach made it impossible to concentrate on participants who fulfilled this criterion.

Device Also the device participants used to complete the experiment was taken into account, laptop (Laptop) being the reference level. Using smartphones or tablets for participation was allowed because too many participants would have been missed who access Facebook or their mail via their smartphone. It was considered unlikely that they access the experiment at a later point again with their laptop at hand.

Handedness Handedness was not expected to have an impact on reaction times, but it was included in the first model for reasons of completeness. The large majority of the participants (88.6 percent) are right-handed (Right), so this was the reference level.

Random intercepts: Participant, StimulusID, ResponsePosition, WordPosition, Letters The participant (i.e., the successively allotted participant ID) functioned as a random intercept in model to make sure that participant-specific effects did not impact on the other predictor estimates. The stimulus ID was added as a random intercept for the same reason. Finally, the position of the stimulus in the task (`ResponsePosition`), the position of the word in the stimulus (`WordPosition`), and

the number of letters for each word (`Letters`) were included as random intercepts in the model in order to account for the impact of those factors on reaction times.

Table 8.5: First language and home languages by group (percentages in brackets)

	SG		HK		IN		control	
first language:								
English	44	(70.97)	13	(36.11)	1	(1.92)	39	(90.70)
Chinese	8	(12.90)	21	(58.33)	0	(0.00)	1	(2.33)
English & Chin.	7	(11.29)	0	(0.00)	0	(0.00)	0	(0.00)
Other	2	(3.23)	0	(0.00)	49[†]	(94.23)	1	(2.33)
English & Other	0	(0.00)	0	(0.00)	2	(3.85)	1	(2.33)
no info	1	(1.61)	2	(5.56)	0	(0.00)	1	(2.33)
sum	62	(100.00)	36	(100.00)	52	(100.00)	43	(100.00)
home languages:								
English	16	(25.81)	10	(27.78)	0	(0.00)	34	(79.07)
Chinese	14	(22.58)	20	(55.56)	0	(0.00)	1	(2.33)
English & Chin.	21	(33.87)	1	(2.78)	0	(0.00)	1	(2.33)
Other	4	(6.45)	2	(5.56)	41	(78.85)	2	(4.65)
English & Other	5	(8.06)	0	(0.00)	9	(17.31)	3	(6.98)
Chinese & Other	0	(0.00)	1	(2.78)	0	(0.00)	0	(0.00)
Engl., Chin. & Oth.	1	(1.61)	0	(0.00)	0	(0.00)	1	(2.33)
no info	1	(1.61)	2	(5.56)	2	(3.85)	1	(2.33)
sum	62	(100.00)	36	(100.00)	52	(100.00)	43	(100.00)

[†]Hindi (19), Telugu (8), Kannada (6), Bengali (3), Tamil (3), Malayalam (2), Tulu (2), Urdu (2), Marathi (2), Gujarati (1), Punjabi (1)

Before we turn to the model, let us have a look at the performance of the different speaker groups in the various conditions tested. Figure 8.2 plots the 95 percent confidence intervals for the reaction times measured by condition and group (compare section 8.3 for details on this way of depicting the reaction times). The reaction times towards the critical regions (target verb or noun plus the three following words) were measured). Conditions V2, V4, N2, and N4 are the "unmarked" conditions. In V1 and V2, a time adverbial precedes the target verb. In N1 and N2, a quantifier precedes the target noun.

A first glimpse at figure 8.2 reveals that all groups show greater variation in their reaction times across the noun conditions than across the verb conditions and that it takes them particularly long to read the critical regions in condition N4. The latter is explicable by the fact that in sentences such as *Bill told us detail about the*

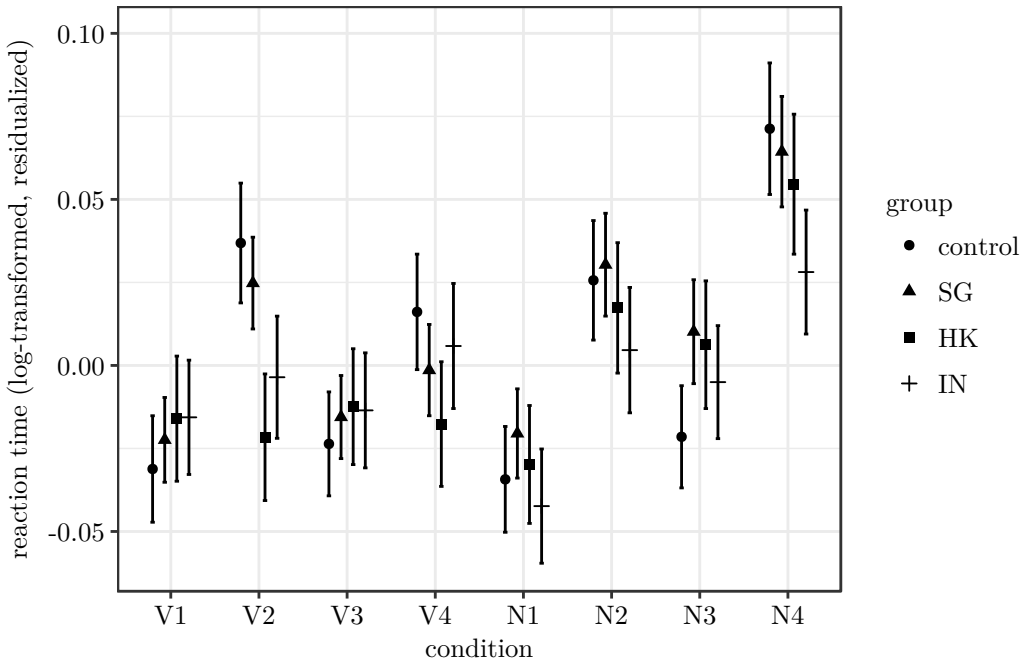

Figure 8.2: 95 percent confidence intervals for reaction times by condition and group

terrible accident he had been involved in (N4) also the article is missing. Recall from table 8.3 that the missing article was necessary to make the four conditions differ in the presence and absence of a quantifier and the plural suffix only.

As expected, all groups read the critical regions in the "unmarked" noun conditions (N2, N4) slower than those in the "marked" noun conditions (N1, N3). For the verb conditions, the tendency is the same except among the HKE speakers, but reading times across groups vary considerably in the "unmarked" conditions (V2 in particular). As pointed out in section 4.3, both speakers of BrE and of AmE served as the control group because their past tense and plural marking systems are assumably stable. In fact, both groups hardly differed in their reaction times to the "marked" and "unmarked" verb and noun conditions (compare figure E.4 in appendix E.4). The only exceptions are conditions V2 and V3, but it is unlikely that the differences observed there are due to variety-internal factors.

Zooming in on the group differences, we clearly see that the control group reads the "unmarked" verb and noun conditions comparatively slowly and the "marked" verb and noun conditions comparatively fast, therefore showing the highest differences in reading speed for "unmarked" versus "marked" conditions. The same trend is observable for the SgE speakers, but only for the verb conditions. The HKE speakers

hardly differ in their reading speed across the verb conditions, but do so for the noun conditions with the expected tendencies ("unmarked" noun conditions are read slower than the "marked" ones). In fact, they read the "unmarked" conditions slightly faster. The IndE speakers differ relatively little in their reading speed across the verb conditions as well and more across the noun conditions. Compared with the other groups, they read the noun conditions fast.

The first model

A first model including all the predictor variables presented above revealed that none of the speaker background variables contributed significantly to the model (`Age`, `AgeEnglish`, `FirstLanguage`, `Sex`, `Education`, `LinguisticsClasses`, `Device`, `Handedness`). Since removing all those predictor variables one by one or considering their interaction with `Condition` brought no noticeable improvement, they were discarded for the final analyses. This is the formula for the first model (`model.spr.pre`):

```
lmer(logRTresidual ~ Condition*Group + Frequency
                       + SpellingChange + Equivalent + Age
                       + FirstLanguage + Sex + Education
                       + LinguisticsClasses + Device + Handedness
                       + (1|Participant) + (1|StimulusID)
                       + (1|ResponsePosition) + (1|WordPosition)
                       + (1|Letters), data=df.spr)
```

Despite their lack of significance, several of the discarded predictor variables show the expected trends. Critical regions containing more frequent words were read faster than critical regions containing less frequent words, and older participants took longer to read the critical conditions than younger ones. Participants with English as a first language and/or with a master's degree or even higher level of education behaved comparatively conservatively, meaning they read the critical conditions more slowly than their counterparts. There was no significant difference in reaction times between participants who used a laptop and those who used a smartphone or a tablet. Table E.3 in appendix E.4 provides the model output for the first model (see below for details on the information provided in that table).

The final model

Having excluded all insignificant predictor variables from the first model, we are left with the following final model (`model.spr`):

```
lmer(logRTresidual ~ Condition*Group + (1|Participant)
                     + (1|StimulusID) + (1|ResponsePosition)
                     + (1|WordPosition) + (1|Letters),
                     data=df.spr)
```

Table 8.6 depicts the model output. It provides the predictor estimates, the standard error, the t-value, and the p-value for the intercept and each model predictor. The main effects of the predictors `Condition` and `Group` are not interpreted because they are involved in significant interactions (cf. Baayen 2012: 166). `Condition` is sum coded, which is why R does not provide the original factor levels. They have been added in square brackets. Condition N1 is not displayed because it served as the reference level. `Group` is dummy coded and the control group represents the reference level for this predictor variable. Note that the coefficient estimates that remained in `model.spr` are close to those in the first model (`model.spr.pre`). Considerable changes in the coefficient estimates from the first to the final model would have been signs of multicollinearity among the predictor variables in the first model.

Table 8.6: Model output (model.spr)

	Estimate	Std. Error	t-value	p-value	
Intercept	0.0065	0.0121	0.5416	0.5881	
Condition					
Condition1[V4]	0.0093	0.0126	0.7392	0.4598	
Condition2[N2]	0.0212	0.0126	1.6881	0.0914	
Condition3[V1]	−0.0371	0.0128	−2.9069	0.0037	**
Condition4[N4]	0.0715	0.0127	5.6385	0.0000	***
Condition5[N3]	−0.0223	0.0127	−1.7619	0.0781	
Condition6[V2]	0.0282	0.0128	2.2048	0.0275	*
Condition7[V3]	−0.0320	0.0127	−2.5287	0.0114	*
Group					
GroupHK	−0.0070	0.0079	−0.8851	0.3761	
GroupSG	0.0040	0.0069	0.5762	0.5645	
GroupIN	−0.0101	0.0072	−1.4025	0.1608	

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition:Group					
Condition1[V4]:GroupHK	−0.0269	0.0124	−2.1591	0.0308	*
Condition2[N2]:GroupHK	−0.0025	0.0123	−0.2064	0.8365	
Condition3[V1]:GroupHK	0.0222	0.0124	1.7884	0.0737	
Condition4[N4]:GroupHK	−0.0112	0.0123	−0.9139	0.3608	
Condition5[N3]:GroupHK	0.0335	0.0122	2.7372	0.0062	**
Condition6[V2]:GroupHK	−0.0453	0.0124	−3.6620	0.0003	***
Condition7[V3]:GroupHK	0.0179	0.0123	1.4468	0.1479	
Condition1[V4]:GroupSG	−0.0222	0.0108	−2.0462	0.0407	*
Condition2[N2]:GroupSG	0.0025	0.0107	0.2351	0.8141	
Condition3[V1]:GroupSG	0.0046	0.0108	0.4257	0.6703	
Condition4[N4]:GroupSG	−0.0120	0.0108	−1.1101	0.2669	
Condition5[N3]:GroupSG	0.0277	0.0108	2.5759	0.0100	**
Condition6[V2]:GroupSG	−0.0146	0.0109	−1.3369	0.1813	
Condition7[V3]:GroupSG	0.0046	0.0109	0.4254	0.6705	
Condition1[V4]:GroupIN	−0.0007	0.0114	−0.0589	0.9530	
Condition2[N2]:GroupIN	−0.0120	0.0113	−1.0568	0.2906	
Condition3[V1]:GroupIN	0.0259	0.0114	2.2736	0.0230	*
Condition4[N4]:GroupIN	−0.0352	0.0114	−3.0885	0.0020	**
Condition5[N3]:GroupIN	0.0240	0.0113	2.1212	0.0339	*
Condition6[V2]:GroupIN	−0.0269	0.0115	−2.3520	0.0187	*
Condition7[V3]:GroupIN	0.0226	0.0114	1.9742	0.0484	*

Note: *p<0.05; **p<0.01; ***p<0.001

Turning to the HKE speakers first, they read the critical regions in the unmarked verb conditions V2 and V4 compared with the overall mean highly significantly and significantly faster than the control group (in line with hypothesis 4). The same trend is observable for conditions N2 and N4, but it is not significant. Additionally, the HKE speakers read the critical regions in condition N3 compared with the overall mean significantly slower than the control group. The SgE and the IndE speakers also read the critical regions in condition N3 significantly slower than the control group (compare figure 8.2). Apart from that, the SgE speakers only read the critical regions in condition V4 compared with the overall mean significantly faster than the control group. For the other conditions, no significant differences are observable.

The IndE speakers read the critical regions in conditions V2 and N4 significantly faster than the control group, but they also read all verb conditions except for V4 significantly slower than the control group.

The findings just described underline the impression gained from a look at the 95 percent confidence intervals for reaction times by condition and group depicted in figure 8.2 above. While there are clear differences in the reading times of the critical regions in the different conditions (particularly with set type N, where the unmarked conditions are read more slowly across groups), the differences between the control group and the target groups are less clear-cut than expected.

In fact, in line with hypothesis 4, the HKE, SgE, and IndE speakers generally read the unmarked verb and noun conditions faster than the control group, which is explicable by a certain familiarity with inflectionally unmarked verbs and nouns (compare chapters 5 and 6). Zooming in, the SgE speakers behave most similarly to the control group, whereas the HKE speakers show the strongest differences, but only for certain conditions. Recall that many of the HKE speakers indicated to have acquired English as their first language. This likely explains why the differences between the HKE speakers and the control group are less pronounced than expected. The IndE speakers differ in comparatively many conditions from the control group but only with marginal significance (except for condition N4). As pointed out in section 6.5, this finding speaks in favor of a certain familiarity of the IndE speakers with bare noun phrases. Surprisingly many of the sampled nouns in ICE-IND lack nominal plural marking (compare section 6.3).

A particular focus lay on the question whether preceding time adverbials and quantifiers influence the perception of inflectionally unmarked verbs and nouns. Figure 8.2 revealed that across groups the critical regions in condition N4 were read comparatively slowly. While the target groups read the critical regions in condition N4 faster than the control group, the difference is not significant. The same is true for condition N2. Interestingly, all three target groups read the critical regions in condition N3 significantly slower than the control group. A plural noun without a preceding article or quantifier is perfectly grammatical, but it might be that speakers of HKE and SgE are to a certain extend used to encountering a bare noun without a preceding quantifier (compare section 6.5 on the impact of substratum transfer on plural omission). As regards the verb conditions, the HKE speakers read the critical regions in condition V2 highly significantly faster than the control group (the

IndE speakers read them significantly faster than the control group) and the critical regions in condition V4 significantly faster (which the SgE speakers did, too).

Model criticism

Lastly, let us turn to model criticism. As mentioned above, a criterion-based approach with slightly modified backward selection was used. For that, only theoretically meaningfully variables were considered in the first model (`model.spr.pre`), and variables that turned out to be non-significant were excluded from the final model (`model.spr`) in a second step. The goodness of fit of the final model compared to the first model was determined by means of AIC. As described before, AIC evaluates the model quality based on the number of predictors the model comprises (cf. Gries 2013: 261). This means that a model with more predictor variables and a resulting higher AIC is of equal explanatory value to a model with fewer predictor variables and a smaller AIC. Interestingly, `model.spr.pre` had an AIC of 17,227.92 and `model.spr` one of 17,383.78, which is why the first model (`model.spr.pre`) should be the preferred model. The description of the first model (subsection **The first model**) revealed that several of the non-significant predictor variables showed the expected trends. Not only were critical regions that contain more frequent words read faster, but older participants read them more slowly than younger ones. Additionally, participants with English as their first language and/or a high level of education behaved conservatively in reading the critical conditions comparatively slowly. It could well be that, overall, the many speaker background variables and stimulus specifics (`Frequency`, `SpellingChange`, `Equivalent`) included in the first model explain the measured reaction times well despite the fact that they do not contribute significantly to the model on an individual basis.

Figure 8.3 plots the residuals of `model.spr` against the values the model predicts. Residuals describe "the difference between the observed and expected [i.e., predicted or fitted] values" (Baayen 2012: 172). Residuals above the horizontal line crossing zero on the y-axis indicate that the predicted values are too low, residuals below this line show that the predicted values are too high, and residuals on the line suggest that observed and expected values are identical.

As figure 8.3 shows, the residuals cluster nicely around the center and no clear patterns are observable. I.e., the residuals are not grouped above or below the horizontal line or subdivided into various clusters. This is a good sign because it indicates that no patterns underlie the measured reaction times that would be explicable by variables the model does not account for. See figures E.5 and E.6 in appendix E.4

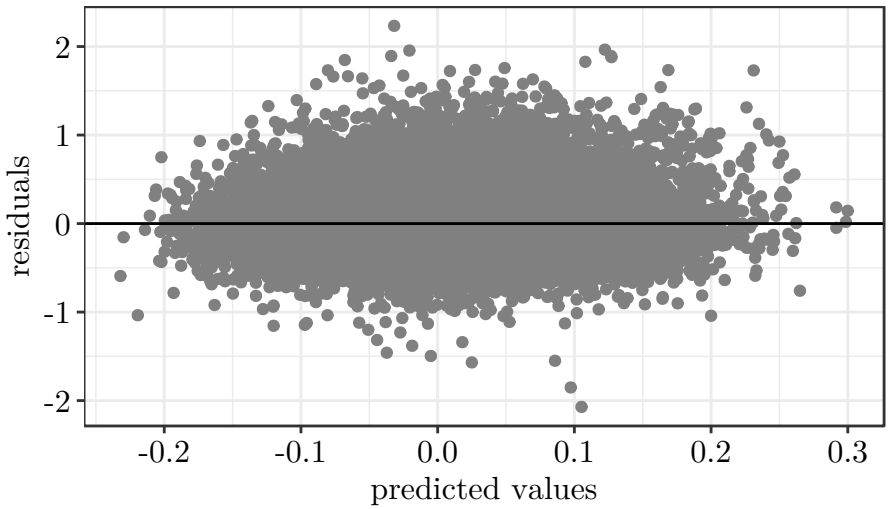

Figure 8.3: Residuals by predicted values (model.spr)

for the residuals by group and condition respectively, where no sub-groupings of the residuals are observable either.

8.4.2 Acceptability judgment task

The main assumption underlying the acceptability judgment task was that speakers of the contact varieties evaluate stimuli containing verbs and nouns that lack inflectional past tense and plural marking more positively than the control group because of their greater familiarity with the features (see hypothesis 5, sections 1.3 and 8.1). This includes inflectionally unmarked verbs and nouns that are preceded by a time adverbial with past time reference and by a quantifier with plural reference. As mentioned in section 8.1, judgments can be influenced by language ideology, which is why participants could have behaved differently than expected.

As with the self-paced reading task, reloaded stimuli were removed prior to the analyses. Additionally, the results of participants who answered 80 percent or less of the comprehension questions correctly were not taken into account. All the stimuli for which the comprehension question had not been answered correctly were not considered either.

As pointed out in the methods section, participants provided their judgments by placing a slider on a continuous scale from “not acceptable at all” to “fully acceptable.” This scale actually represented 100 data points from zero (“not acceptable at all”) to 100 (“fully acceptable”), so each position of the slider on the scale rep-

resented a numeric value that was used for the analyses. The acceptability ratings were z-standardized by participant in order to rule out variation in using the scale across participants (e.g., Sprouse et al. 2013: 228; Staum Casasanto et al. 2010: 226). By means of z-standardization, z-scores were computed "which indicate how many standard deviations each [judgment score] deviates from the mean [of all judgment scores]" (Gries 2013: 122). The mean of the resulting z-scores was 0, the standard deviation 1. Consequently, the judgments of single participants were normalized based on all judgments provided across participants.

Preparing the model

The dependent variable was the z-standardized numeric acceptability judgment. The predictor variables were the same as for the self-paced reading task, with the exception of the following deviations. This time, the distractor stimuli were included in the model. For the self-paced reading task, only set types V and N had been taken into account. As to `Frequency`, the frequency of the critical word in each stimulus was accounted for. Not of interest were the predictor variables `Equivalent`, `Device`, `Handedness`, as well as the random intercepts `Letters` and `WordPosition` because those factors are unlikely to impact on the judgments. `SpellingChange` was not considered either because it is only a meaningful characteristic for set types V and N and not for the distractor stimuli.

Before we turn to the model, let us consider the judgments provided by the different speaker groups for the various conditions tested. Figures 8.4 and 8.5 plot the 95 percent confidence intervals for the judgment scores for set types V, N, and D by speaker group.

The more positively a stimulus was evaluated, the higher its judgment score is. In conditions V2, V4, N2, and N4, the target verb or noun is not marked for past tense and plural, in the remaining verb and noun conditions it is. Conditions V1 and V2 additionally contain a preceding time adverbial, conditions N1 and N2 a preceding quantifier. Among the distractor stimuli, D5 and D6 contain instances of object pronoun drop, D7 and D7 instances of subject pronoun drop, D9 and D10 are examples of conjunction doubling, and D11 and D12 have no non-standard grammatical feature (see table 8.4). D6, D8, D10, D12 additionally contain semantically awkward material. Speakers of BrE and speakers of AmE served as one single control group because both speaker groups hardly differ in their judgments (compare figure E.7 in appendix E.4). The only exceptions here are conditions D7 and D8, but again the differences therein are unlikely due to variety-internal factors.

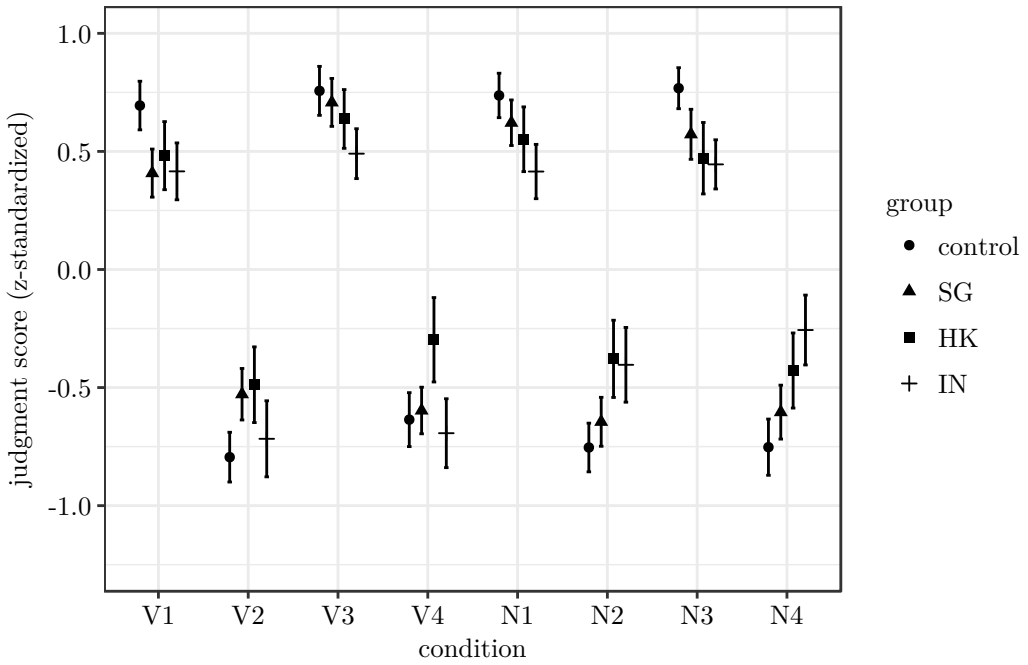

Figure 8.4: 95 percent confidence intervals for judgment scores by condition and group (set types V and N)

Figure 8.4 shows that the "unmarked" conditions (V2, V4, N2, N4) were evaluated much more negatively than the "marked" conditions by all groups. The greatest difference in the judgments of "unmarked" versus "marked" conditions is observable in the control group. Compared with the target groups, the control group accepted the "unmarked" conditions relatively little and the "marked" conditions relatively much. The HKE speakers evaluated the "unmarked" and "marked" conditions most alike. Regarding the distractor stimuli, the stimuli containing semantically awkward material were evaluated more negatively than those without awkward semantics (the only exception being HKE speakers who evaluated D9 on average more negatively than D10). This shows that the participants paid attention to sentence contents. Stimuli with subject pronoun drop (D7, D8) received nearly as high ratings as stimuli without a non-standard grammatical feature (D11, D12), which is a clear sign that sentences such as *Have no clue which gym to pick, but I have to do more for my fitness* (see table 8.4) were considered colloquial but acceptable.

The first model

A first model with all the predictor variables (except for `SpellingChange`, `Device`, `Equivalent`, `Handedness`, and the random intercepts `Letters` and `WordPosition`)

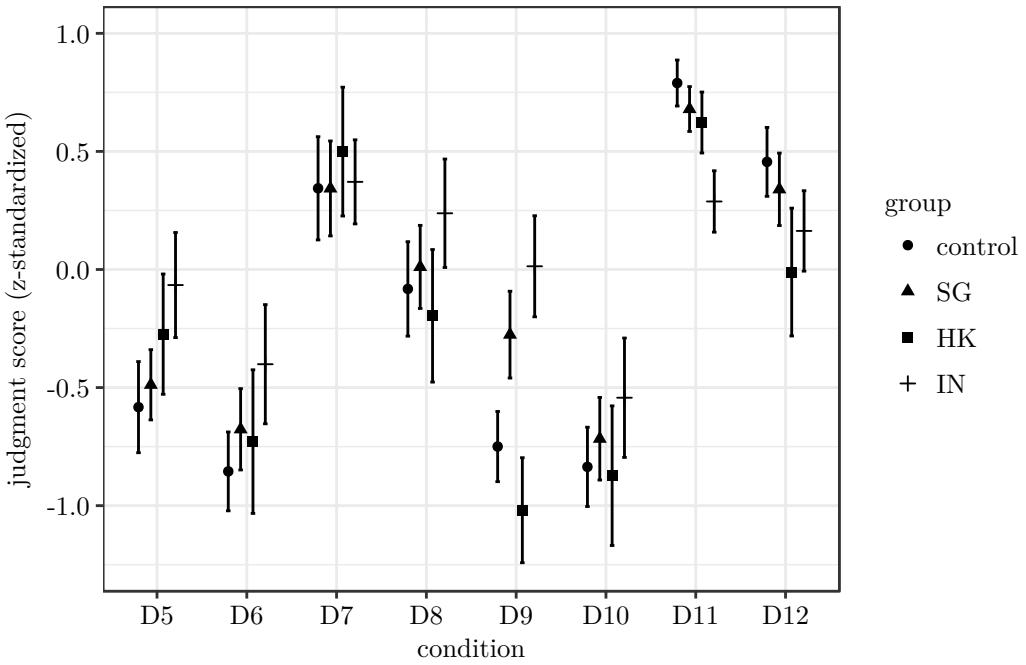

Figure 8.5: 95 percent confidence intervals for judgment scores by condition and group (set type D)

revealed that, as with the self-paced reading task, none of the speaker background variables contributed significantly to the model. `Frequency` proved to have a significant effect insofar as stimuli were evaluated more positively the more frequent the critical word in the stimulus is (compare table E.4 in appendix E.4). This is the formula underlying the first model (`model.ajt.pre`):

```
lmer(JudgmentScore ~ Condition*Group + Frequency + Age
                     + FirstLanguage + Sex + Education
                     + LinguisticsClasses + (1|Participant)
                     + (1|StimulusID) + (1|ResponsePosition),
                     data=df.ajt)
```

Although several of the speaker background variables turned out to be insignificant predictors, they showed tendencies worth mentioning. For evidence, the interested reader is referred to table E.4 (appendix E.4). Participants with Chinese or English and Chinese as their first language(s) evaluated the stimuli more positively than participants with English as their first language (reference level). Participants who have not received a bachelor's degree (yet) and participants who have not attended linguistics classes evaluated the stimuli more positively than those with a master's

degree or higher and those who have attended linguistics classes. This goes in line with the assumption that a higher level of education and background knowledge in linguistics prompt more conservative, i.e., critical, reactions towards the stimuli.

The final model

After excluding all insignificant predictor variables (all speaker background variables in this case) from the first model, the final model (`model.ajt`) looks as follows:

```
lmer(JudgmentScore ~ Condition*Group + Frequency
                     + (1|Participant) + (1|StimulusID)
                     + (1|ResponsePosition), data=df.ajt)
```

Table 8.7 shows the model output, providing the predictor estimates, the standard error, the t-value, and the p-value for the intercept as well as all model predictors. Again, the main effects of `Condition` and `Group` cannot be interpreted because they are involved in meaningful interactions. The factor levels for `Condition` (sum coded) are provided in square brackets. As with the self-paced reading task, the coefficient estimates that remained in the final model (`model.ajt`) are similar to those in the first model (`model.ajt.pre`). Considerably different coefficient estimates in both models would have been signs of multicollinearity in the first model.

Table 8.7: Model output (model.ajt)

	Estimate	Std. Error	t-value	p-value	
Intercept	−0.4712	0.1421	−3.3158	0.0009	***
Condition					
Condition1[V4]	−0.5023	0.0831	−6.0459	0.0000	***
Condition2[D7]	0.2805	0.1882	1.4909	0.1360	
Condition3[D8]	0.1555	0.1848	0.8412	0.4002	
Condition4[D5]	−0.5128	0.1783	−2.8764	0.0040	**
Condition5[D12]	0.6405	0.1325	4.8351	0.0000	***
Condition6[N1]	0.8392	0.0811	10.3419	0.0000	***
Condition7[D9]	−0.7681	0.1822	−4.2156	0.0000	***
Condition8[D6]	−0.8675	0.1823	−4.7584	0.0000	***
Condition9[V3]	0.8905	0.0822	10.8292	0.0000	***

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition10[V2]	−0.6647	0.0841	−7.9076	0.0000	***
Condition11[N3]	0.8764	0.0812	10.7990	0.0000	***
Condition12[D10]	−0.8549	0.1858	−4.6016	0.0000	***
Condition13[N4]	−0.6469	0.0816	−7.9238	0.0000	***
Condition14[V1]	0.8255	0.0825	10.0017	0.0000	***
Condition15[N2]	−0.6539	0.0810	−8.0731	0.0000	***
Group					
GroupHK	0.0031	0.0310	0.0997	0.9206	
GroupSG	0.0383	0.0256	1.4931	0.1354	
GroupIN	0.0792	0.0273	2.9039	0.0037	
Stimulus specifics					
Frequency	0.0370	0.0138	2.6836	0.0073	**
Condition:Group					
Condition1[V4]:GroupHK	0.3296	0.0961	3.4284	0.0006	***
Condition2[D7]:GroupHK	0.1574	0.1662	0.9467	0.3438	
Condition3[D8]:GroupHK	−0.1079	0.1469	−0.7346	0.4626	
Condition4[D5]:GroupHK	0.3049	0.1441	2.1161	0.0343	*
Condition5[D12]:GroupHK	−0.4568	0.1064	−4.2922	0.0000	***
Condition6[N1]:GroupHK	−0.1907	0.0942	−2.0240	0.0430	*
Condition7[D9]:GroupHK	−0.2762	0.1438	−1.9212	0.0547	.
Condition8[D6]:GroupHK	0.1092	0.1490	0.7331	0.4635	
Condition9[V3]:GroupHK	−0.1219	0.0953	−1.2799	0.2006	
Condition10[V2]:GroupHK	0.3008	0.0965	3.1186	0.0018	**
Condition11[N3]:GroupHK	−0.2990	0.0948	−3.1534	0.0016	**
Condition12[D10]:GroupHK	−0.0401	0.1521	−0.2637	0.7920	
Condition13[N4]:GroupHK	0.3249	0.0952	3.4117	0.0006	***
Condition14[V1]:GroupHK	−0.2288	0.0959	−2.3867	0.0170	*
Condition15[N2]:GroupHK	0.3640	0.0948	3.8390	0.0001	***
Condition1[V4]:GroupSG	−0.0044	0.0793	−0.0550	0.9561	
Condition2[D7]:GroupSG	−0.0561	0.1397	−0.4018	0.6879	
Condition3[D8]:GroupSG	0.0513	0.1200	0.4277	0.6689	
Condition4[D5]:GroupSG	0.0565	0.1195	0.4727	0.6365	

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition5[D12]:GroupSG	−0.1467	0.0895	−1.6386	0.1013	
Condition6[N1]:GroupSG	−0.1500	0.0783	−1.9156	0.0554	.
Condition7[D9]:GroupSG	0.4386	0.1173	3.7382	0.0002	***
Condition8[D6]:GroupSG	0.1408	0.1208	1.1661	0.2436	
Condition9[V3]:GroupSG	−0.0903	0.0787	−1.1476	0.2511	
Condition10[V2]:GroupSG	0.2274	0.0808	2.8151	0.0049	**
Condition11[N3]:GroupSG	−0.2441	0.0784	−3.1143	0.0018	**
Condition12[D10]:GroupSG	0.0810	0.1265	0.6405	0.5219	
Condition13[N4]:GroupSG	0.1024	0.0789	1.2982	0.1942	
Condition14[V1]:GroupSG	−0.3307	0.0793	−4.1717	0.0000	***
Condition15[N2]:GroupSG	0.0721	0.0781	0.9232	0.3559	
Condition1[V4]:GroupIN	−0.1535	0.0867	−1.7691	0.0769	
Condition2[D7]:GroupIN	−0.0075	0.1425	−0.0524	0.9582	
Condition3[D8]:GroupIN	0.2362	0.1277	1.8498	0.0643	
Condition4[D5]:GroupIN	0.4384	0.1261	3.4766	0.0005	***
Condition5[D12]:GroupIN	−0.3670	0.0955	−3.8440	0.0001	***
Condition6[N1]:GroupIN	−0.3980	0.0846	−4.7052	0.0000	***
Condition7[D9]:GroupIN	0.6824	0.1267	5.3856	0.0000	***
Condition8[D6]:GroupIN	0.3706	0.1299	2.8540	0.0043	**
Condition9[V3]:GroupIN	−0.3390	0.0840	−4.0366	0.0001	***
Condition10[V2]:GroupIN	−0.0065	0.0865	−0.0746	0.9405	
Condition11[N3]:GroupIN	−0.4048	0.0836	−4.8442	0.0000	***
Condition12[D10]:GroupIN	0.2178	0.1335	1.6313	0.1028	
Condition13[N4]:GroupIN	0.4054	0.0856	4.7361	0.0000	***
Condition14[V1]:GroupIN	−0.3742	0.0853	−4.3857	0.0000	***
Condition15[N2]:GroupIN	0.2752	0.0847	3.2482	0.0012	**

Note: *p<0.05; **p<0.01; ***p<0.001

As in the first model, `Frequency` proved to have a significant effect insofar as stimuli were evaluated more positively the more frequent the critical word in the stimulus is. In an extra step, the interaction of `Frequency` with `Condition` was accounted for to see whether this is true for certain conditions in particular. Since no significant interactions of `Frequency` with `Condition` were observable, the more

general finding applies: Across conditions a higher frequency of the critical words goes in line with more positive evaluations.

Turning to the interaction of `Group` with `Condition`, a first glance at table 8.7 reveals that the judgments provided by the SgE speakers deviate least from the judgments provided by the control group. As pointed out above, judgments can be influenced by language ideology and are therefore not easily predictable. As we saw in figure 8.4, the control group accepted the "unmarked" conditions comparatively little and the "marked" conditions comparatively much. The model findings indicate that the SgE speakers adopted this conservative stance. Of all critical conditions, the SgE speakers only evaluated V2 significantly more positively than the control group (in line with hypothesis 5). They also evaluated V4, N2, and N4 more positively, but not significantly. This is in stark contrast to the HKE speakers, who evaluated V2 significantly and V4, N2, and N4 highly significantly more positively than the control group. The IndE speakers are in between. While they evaluated V2 and V4 more negatively than the control group (no significant effect though), they approved of N2 and N4 highly significantly more than the BrE and AmE speakers tested. As pointed out before, the comparatively high approval of bare noun phrases (N4) among the IndE participants invites future research on that topic.

As to the distractor stimuli, the SgE again behaved similarly to the control group and only evaluated condition D9 (conjunction doubling, no semantically awkward lexical material) highly significantly more positively than the control group. The HKE speakers evaluated condition D5 (object pronoun drop, no semantically awkward lexical material) significantly more positively and condition D12 (no feature, semantically awkward lexical material) highly significantly more negatively than the control group. In doing so, they behaved more standard-like than the IndE speakers, who approved of condition D5 significantly more and of D6 (object pronoun drop, semantically awkward lexical material) and D9 highly significantly more than the control group. Like the HKE speakers, the IndE speakers evaluated D12 highly significantly more negatively than the control group.

Because of the high degree of variation in the judgments of the distractor stimuli, a further trimmed model was fitted that focused on the verb and noun conditions exclusively. There is no room to elaborate on this model in detail here, but it is sufficient to say that the verb and noun conditions show the same tendencies, irrespective of whether the distractor stimuli are accounted for or not. This is important to point out because the overall mean judgment score in `model.ajt` includes the judgments

of the distractor stimuli. A strong impact of the distractor judgments on the model findings can therefore be ruled out (compare table E.5 in appendix E.4 for the model output of the respective model called `model.ajt.vn`).

Model criticism

Let us evaluate the final model (`model.ajt`) in a last step. The goodness of fit of the final model in contrast with the first model was determined by means of AIC. The AIC of the final model (19579.8) is slightly higher than that of the first model (19382.25). Recall that AIC shows the model quality based on the number of predictors the model has. In the discussion of `model.ajt.pre`, we learned that several of the non-significant predictor variables show the expected trends. Participants with Chinese or English and Chinese, participants with a low level of education, and those who have no background knowledge in linguistics evaluated the stimuli across conditions comparatively positively. Again, it could be that while those background variables do not contribute significantly to the model, they explain the judgments provided quite well.

Figure 8.6 plots the residuals of `model.ajt` against the judgment scores the model predicts. The residuals cluster along the horizontal line crossing zero on the y-axis, and no clear patterns are observable. Figures E.8 and E.9 in appendix E.4 depict the residuals subdivided by group and condition and show no clear patterns either, which signals that no variables were missed when fitting the model that would have explained the measured judgment scores.

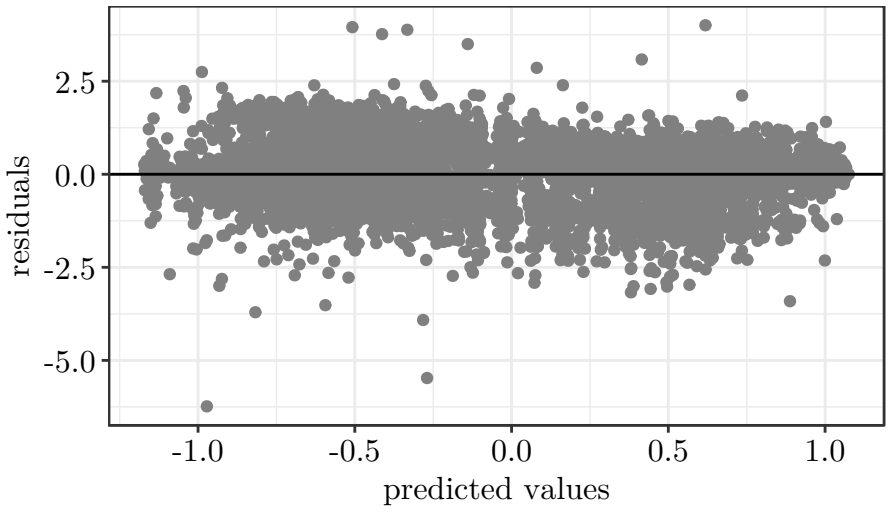

Figure 8.6: Residuals by predicted values (model.ajt)

8.5 Discussion

The results of both the self-paced reading task and the acceptability judgment task turned out as expected. All target groups as well as the control group read the conditions containing inflectionally unmarked verbs and nouns slower than their marked counterparts and evaluated them more negatively. Reading times in the "unmarked" versus "marked" conditions differ most strongly in the control group and least so among the HKE speakers; particularly so in the verb conditions. The same is true for the judgments provided. These findings underline the results of the corpus studies on omission of verbal past tense marking (chapter 5) and nominal plural marking (chapter 6): Omission rates turned out to be surprisingly low in the Asian contact varieties but considerably higher than in the lexifier BrE. The impression gained from the experiment is that the target groups, and the HKE speakers in particular, are more used to lack of inflectional marking than the control group.

Regarding the impact of preceding time adverbials and quantifiers on the reading times and judgments measured, it is worth pointing out that all groups had difficulties with condition N4, the "unmarked" noun condition without a preceding quantifier; in both tasks. Compared with the respective verb condition V4 and the "unmarked" noun condition with a preceding quantifier (N2), participants seemed to struggle with stimuli such as *Bill told us detail about the terrible accident he had been involved in* (condition N4). While it is arguably the less likely solution to fix the stimulus, an article preceding the bare noun could be missing. This issue would invite further testing. The results obtained here are clear proof that, given the experiment is carefully planned and the data are thoroughly analyzed, web-based testing is a valid alternative to on-site testing.

9 Determinants of simplification: A general discussion

This chapter evaluates the explanatory power of the determinants of simplification discussed in the previous chapters based on the corpus and experiment findings. The overarching aim of the book was to figure out how far frequencies of use impact omission and regularization and to which degree substratum transfer and institutionalization function as constraints on frequency. While brief summaries were already presented with each study, the idea here is to take a broader view and connect the findings from a bird's eye perspective. Additionally, the learner character and universality of the phenomena investigated are elaborated on.

Usage frequency

Regarding frequencies of use, lemma token frequencies served as the main frequency measure. In the corpus study on omission of verbal past tense marking, type frequencies were additionally accounted for. This was due to the fact that the corpus study on past tense omission contrasted past tense omission in regular verbs with omission in irregular verbs. Most regular verbs have lower token frequencies individually than irregular verbs, i.e., they occur comparatively infrequently. At the same time, they have a much higher type frequency as a class than irregular verbs due to the fact that the majority of English verbs have adopted regular /t,d/ affixation throughout the course of history.

Different frequency assumptions underlay the investigation of the omission and regularization phenomena considered. For cases in which omission of inflectional marking is morphologically conditioned, infrequent forms were assumed to be affected by omission more (or first) and frequent forms less (or later). This was expected to be the case for regular as well as irregular forms (see hypothesis 1a, section 1.3). The focus was on verb and noun lemmata whose inflectional past tense and plural marking is not prone to consonant cluster reduction. This way phonetically motivated omission was ruled out. With regularization, it was assumed that irregular forms are affected by regularization more (or first) and frequent irregular forms less (or later; see hypothesis 1b, section 1.3).

Across features and varieties, it was striking how low omission and regularization rates turned out to be once interfering factors like the phonetic environment had been ruled out and potential hits had been checked carefully for viability. For regular verbs in ICE-SIN in particular, it was observed that omission is mainly restricted to environments in which the [t,d] suffix is prone to consonant cluster reduction when both the sounds preceding and following the suffix are accounted for. Omission of both past tense and plural marking proved to be higher in ICE-HK than in ICE-SIN and ICE-IND.

The key findings from the corpus analyses on omission and regularization are twofold. Firstly, phonetic reduction seems to be crucial for lack of verbal past tense marking in SgE and HKE. This finding is worth testing further by means of corpora of spoken language, including the respective sound recordings—particularly because omission of verbal past tense marking is a feature commonly described for SgE and not exclusively for consonant cluster environments. The corpus results presented in chapter 5 leave the impression that past tense omission in SgE is to a large extent phonologically conditioned though.

Secondly, not many highly frequent forms are prone to omission and regularization. It is worth pointing out that verbs with irregular past tense marking are comparatively little affected by omission, even when irregular verbs of similar frequency to regular verbs were accounted for (see table 5.4 in section 5.4). The differences in the mental representation of regular forms and irregular forms (cf. Croft & Cruse 2004: 292–296) and the salience of irregular forms are likely to account for that. The generally low regularization rates observed further support this finding.

Substratum transfer

Turning to substratum transfer next, it was assumed that substratum transfer functions as a constraint on frequency effects in case the simplification phenomena considered predominate in SgE and HKE irrespective of lemma frequency (see hypothesis 2, section 1.3). Obviously, the observed patterns of simplification need to be explicable by common substrate influence for the argument to hold.

As regards past tense omission, we clearly saw two effects of substratum transfer on omission rates. Regarding SgE in particular, substratum-driven consonant cluster reduction is likely an explanation for the differences in omission rates in consonant versus vowel environments (see section 5.3). HKE shows the same pattern, but the difference is less pronounced. Substratum transfer functions as a constraint on frequency effects insofar as substratum-driven consonant cluster reduction affects

SgE and HKE verbs, irrespective of their frequency of occurrence. I.e., the effects of substratum transfer counteract expected frequency effects.

A second effect is visible when omission rates in perfective and imperfective contexts are contrasted. Here we learned that across varieties omission rates prevail in imperfective contexts, which was expected based on substrate influence from Mandarin and other Chinese dialects, as well as Malay and Hindi (compare section 5.5). As to omission of nominal plural marking, we saw that the existence of pre-nominal elements, such as numerals that indicate plural reference in Chinese dialects like Mandarin and Cantonese, is one possible explanation why nouns themselves remain unmarked for plural and why the plural reference is indicated by premodifying numerals or quantifiers instead or not at all. In HKE, for instance, comparatively high plural omission rates were observed after quantifiers, which is explicable by the fact that pre-nominal elements like numerals are used to mark plural in Cantonese. Concerning the use of uncountable nouns as countable nouns, it is worth mentioning that the lack of consistent distinctions between countable and uncountable nouns in English leads to confusion among Chinese mother tongue learners of English (cf. Liu et al. 2006: 137). Further noun types such as zero plurals and collectives pose additional difficulties. The observed differences in omission rates in imperfective versus in perfective contexts, the impact of preceding numerals and quantifiers, and the lack of consistent distinctions between countable and uncountable nouns impact verbal past tense and nominal plural marking without usage frequencies of the respective verbs and nouns playing a role.

One clear advantage of the perception experiment compared with the corpus data is that a detailed speaker background questionnaire preceding the experiment provided information about the age, sex, first language, and level of education of participants, among other things. Participants who speak English as their first language and participants who have reached the highest level of education (master's degree or higher) took comparatively long to read the critical regions, whether the target verbs and nouns were marked or unmarked. They also evaluated the stimuli more negatively across conditions. The fact that L1 speakers of English reacted more conservatively in both tasks is a clear sign that L1 speakers of Chinese or of one of the other languages mentioned (many of them Indian languages) were less critical with the stimuli. The more language background information is given, the more straightforward claims on effects of substratum transfer can be made.

Institutionalization

It was hypothesized that institutionalization functions as a constraint on frequency effects in case omission and regularization patterns are particularly stable in SgE irrespective of lemma token frequency (see hypothesis 3, section 1.3). SgE is most institutionalized among the contact varieties of interest, so comparatively stable simplification patterns were expected for this variety.

The corpus findings on verbal past tense and nominal plural marking clearly showed that omission is less patterned in ICE-HK than in ICE-SIN and ICE-IND. Past tense omission in ICE-SIN turned out to be mainly restricted to environments where the regular past tense suffix is prone to consonant cluster reduction, irrespective of frequencies of use. This is why it was concluded that institutionalization does not function as a constraint on frequency in SgE. In ICE-HK, also *-ed* suffixes in vowel environments as well as irregular verbs are affected by omission. In general, past tense and plural omission rates along the frequency cline showed comparatively little systematicity in ICE-HK (in contrast with ICE-SIN). Apart from that, Biewer's (2015: 173) observation that L2 learners struggle with bound morphemes likely explains why speakers of HKE, arguably L2 speakers of English, are comparatively likely to omit the past tense and plural suffix. Regarding regularization, regularization rates turned out to be very low across the relevant varieties (irrespective of degrees of institutionalization). Neither does SgE pattern with the control BrE, nor is HKE particularly prone to regularization.

Similarly, the experiment revealed that the HKE speakers differ most strongly from the control group in their reading times and acceptability judgments, which ties in with the little patterned omission rates in ICE-HK. The results of the SgE speakers turned out to be most similar to those of the control group, and the IndE speakers patterned in between. The SgE group seemed to be much more aware of omission of verbal past tense and nominal plural marking than the other target group speakers (HK and IN), which is not only apparent from their judgment scores but also from their reading times. It is likely that this metapragmatic assessment also affects language production of SgE speakers—an assumption not tested here.

Considering frequencies of use, substratum transfer, and institutionalization as determinants of omission and regularization clearly proved to be revealing. As the main findings summarized above show, substratum transfer and institutionalization constrain frequency effects whenever they counteract expected frequency effects. The findings imply that, given the conditions substratum transfer and learner behavior

provide, language patterns of speakers of little institutionalized varieties are comparatively little systematic. This even extends to highly frequent (irregular) forms. As long as a sufficient amount of language data is available to meaningfully account for frequency effects (chapter 10), a frequency-based approach to World Englishes thus offers valuable insights that help explain linguistic patterns in contact varieties.

Omission and regularization as learner phenomena

One of the distinctions repeatedly referred to in the corpus analyses chapters is Van Rooy's (2011) notion of errors versus conventionalized innovations. Grammatical stability and acceptability are clear signs of the latter, and while errors are observable among both speakers of Foreign Language Englishes and speakers of New Englishes, conventionalized innovations prevail in the New Englishes (compare section 2.4). The distinction between errors and conventionalized innovations directly ties in with recent discussions about differences (or lack thereof) between ESL and EFL varieties, where the crucial point of debate is whether deviations from standard usages should be referred to as errors in the case of EFL varieties when they are labeled innovations in ESL usage. This "paradigm gap" was already criticized by K. K. Sridhar & S. N. Sridhar (1986), resulting in a "plea for an integrated approach" (Hundt & Mukherjee 2011: 1). The corpus studies and the experiment presented here raise the question whether sporadic and unsystematic omission and regularization patterns should be considered erroneous, innovative, or something else. Recall Edward's (2014) reasoning that ESL and EFL varieties "share a common acquisitional starting point, which results in similar strategies such as transfer, redundancy and regularisation" (173). The fact that transfer, (avoidance of) redundancy, and regularization are strategies that are adopted by both ESL and EFL speakers makes it difficult to speak of innovations in one case and of errors in the other case.

Among the features discussed in this book, the use of uncountable nouns as countable nouns seems to be the most likely candidate for a learner feature. A considerable variety of noun types exist that differ as to whether they have a formal singular plural contrast, a meaning difference between singular and plural forms, and as to whether they take singular or plural concord (compare section 7.2.1). All noun types, except for countable nouns with regular plural marking, constitute minority types, meaning their respective uses have to be memorized.

Of the varieties considered in this book, HKE is the one whose variety status has been a matter of debate in the (recent) past (cf. Schneider 2007: 137). At the same time, the past tense and plural omission rates in HKE turned out to be comparatively

high and unsystematic. Given Hong Kong's status as a former British colony, HKE is typically not referred to as a learner variety in the literature. Bolton (2002b: 44–47), for instance, refrains from speaking of a distinct variety, and Evans (2014: 592) argues that the conditions for the emergence of a nativized variety are not given in Hong Kong because of limited use of English in the private domain. These observations obviously raise the question where to draw the line between the acquisition of English in EFL settings such as, say, China and ESL settings such as Hong Kong. While English is one of the official languages in Hong Kong, only a very small minority of the population reported in 2006, 2011, and 2016 that English is the language they usually use (see table 3.4, section 3.3.1). Speakers of English from Mainland China are as likely as HKE speakers to transfer lack of inflectional marking for past tense or plural from the Chinese dialects they speak. A comparative study of the two speaker groups on past tense and plural omission would be worth conducting. ICLE contains argumentative essays written by Chinese EFL learners of English and would be a good starting point. Even better would be recordings of spontaneous speech.

Omission and regularization as universal phenomena?

While omission of verbal past tense and nominal plural marking is nearly non-existent in ICE-GB, regularized irregular verbs and uncountable nouns attaching the regular plural suffix *-s* occur as sporadically in GloWbE GB (and GloWbE US in the latter case) as in GloWbE SG, GloWbE HK, and GloWbE IN. This raises the question whether the regularization phenomena discussed here are universal phenomena that occur across varieties and variety types. As mentioned in section 1.1, Chambers (2004) defines vernacular universals as "a small number of phonological and grammatical processes [that] recur in vernaculars wherever they are spoken" (128).

Section 2.1 elaborated on simplification in eWAVE. It accounted for top L2, top Asian, and L2-simple features in CSE, HKE, and IndE and additionally listed the L2-simple features that occur more frequently in Asia than globally. The eWAVE ratings revealed that the simplification features of interest are most salient in HKE, followed by CSE and IndE, which mirrors the degree to which the phenomena were observable in the corpus data: omission and regularization turned out to be most prevalent in HKE and least prevalent in IndE.

Table 9.1 lists the eWAVE ratings for the simplification features of interest in English dialects in the North, Southwest, and Southeast of England. Those are the regions that represent BrE, the lexifier for the Asian varieties considered, in eWAVE. Additionally, the table shows for each feature whether it is a top L2 feature (cf.

Kortmann & Wolk 2012) or an L2-simple feature (cf. Szmrecsanyi & Kortmann 2009: 69–71; Kortmann & Szmrecsanyi 2009: 274). With the exception of missing information for a few features in dialects in the Southwest of England, lack of plural marking (features 57 and 58), zero past tense forms (feature 132), and the use of plural for StE singular mass nouns (feature 55) are reported for none of the regions. The regularization of irregular verb paradigms (feature 128) received B-ratings in North and Southwest English dialects, and an absence of plural marking only after quantifiers (feature 56) is reported for all three dialect regions.

Table 9.1: The features of interest in English dialects in the North (N-E), Southwest (SW-E), and Southeast (SE-E) of England in eWAVE (Kortmann & Lunkenheimer 2013)

feature no.	ratings† for English dialects in			top L2	L2-simple
	N-E	SW-E	SE-E		
55	D	?	D	✓	
56	B	B	B		(✓)
57	D	D	D		
58	D	D	D		
128	B	B	C		✓
132	D	?	D		✓

feature no.	feature
55	different count/mass noun distinctions: use of plural for StE singular
56	absence of plural marking only after quantifiers
57	plural marking generally optional: for nouns with human referents
58	plural marking gen. optional: for nouns with non-human referents
128	regularization of irregular verb paradigms
132	zero past tense forms of regular verbs

†ratings in eWAVE: A - feature is pervasive or obligatory; B - feature is neither pervasive nor extremely rare; C - feature exists, but is extremely rare; D - attested absence of feature; X - feature is not applicable; ? - no information on feature is available

It would be worth comparing the Asian varieties analyzed here in more detail with dialects from the British Isles in order to see whether omission and regularization occur in English dialects to a certain extent as well; although the features are not salient, as the eWAVE ratings presented in table 9.1 show. Detailed analyses of the features that account for all aspects considered in the corpus studies and that provide comparability with the findings in ICE cannot be provided here. The inter-

active database of the *Freiburg Corpus of English Dialects* (FRED; cf. Hernández 2006; Universitätsbibliothek Freiburg 2018) is a handy tool for conducting respective studies and would allow for comparative analyses. FRED is described as "a monolingual spoken-language dialect corpus" (Hernández 2006: 1) that comprises 372 texts with a total of about 2.5 million words representing 300 hours of speech recordings (ibid.: 2). The recorded conversations stem from the years 1968 to 2000 and from the 1970s and 1980s in particular. The 432 informants are from nine dialect areas (Isle of Man, Hebrides, Midlands, North, Scottish Highlands, Scottish Lowlands, Southeast, Southwest, and Wales) that are further subdivided into counties. 63.7 percent of the informants are male, 30.6 percent female (for the rest, the sex is not known), and the age range is six to 102 years, 75 being the mean age. The recently released interactive database enables conducting full text searches on the basis of FRED or FRED-S, a subset of about 1 million words from 121 interviews that have been transcribed orthographically and that cover five dialect areas (Midlands, North, Scottish Lowlands, Southeast, and Southwest). Only FRED-S is fully accessible online, i.e., the audio files, plain texts, and tagged text files are available for download. As part of the full text searches, the age and sex distribution for the findings is provided. The findings can be filtered for area, county, and location.

In section 2.1, Mesthrie's (2012) feature density account was introduced, which deals with the distribution of the eWAVE features across world regions (FD world) and in Asia (FD Asia), among other things. Mesthrie (2012) uses "Edgar Schneider's criterion of 80% as a cut-off point to indicate feature density, i.e., that a particular feature occurs in 80% of the Asian varieties categorized in WAVE" (786). A feature density of 80 percent or more is considered high. Table 9.2 depicts the feature densities for Asia and the world, their difference, and the feature area each feature belongs to. For the feature titles, consult the lower half of table 9.1.

Table 9.2: Feature densities for Asia and the world for the features of interest in eWAVE (cf. Mesthrie 2012)

feature no.	FD Asia (%)	FD world (%)	difference (%)	feature area
55	100.0	54.1	45.9	noun phrase
56	28.6	43.2	-14.7	noun phrase
57	71.4	40.5	30.9	noun phrase
58	71.4	41.9	29.5	noun phrase
128	71.4	63.5	7.9	verb morphology
132	71.4	59.5	11.9	verb morphology

A comparison of the feature densities for Asia and the world reveals that all the features except for feature 56 have a higher feature density in Asia than across world regions. The differences are particularly strong for features 55, 57, and 58. Based on the assumption that universal features have a high global feature density, it can be concluded that none of the simplification features analyzed in this book are a potential candidate for a universal.

10 Concluding remarks and implications

To conclude, the idea to consider frequencies of use, substratum transfer, and institutionalization as factors that impact omission and regularization proved to be revealing. While substratum transfer and institutionalization have been the subject of investigation in numerous studies on World Englishes, in-depth usage-based accounts are rare.

Usage-based research: Methodological implications

Considering usage frequencies had clear methodological implications and ICE and GloWbE proved to be the most suitable corpora available for the purpose of the analyses conducted (see section 4.1). For the corpus studies on omission in ICE, all spoken sections and pseudo-random samples of lemmata were considered to be left with high enough frequencies of occurrence. Rather than exploring patterns of past tense and plural marking by going through the corpus files verb by verb and noun by noun, the focus was strictly on missing past tense and plural marking in the verb and noun samples. Regarding GloWbE, apart from working with samples of verb and noun lemmata, it was necessary to limit the number of hits for unmarked verbs and nouns to samples (200 hits in GloWbE SG and GloWbE HK and 400 hits in GloWbE IN) to be able to manually identify bare verb and noun forms that lack past tense and plural marking, respectively. In contrast, regularized irregular verbs and uncountable nouns with the regular plural suffix could be directly retrieved from GloWbE. These observations show that commonly used corpora in World Englishes research like ICE and GloWbE are suitable for investigating the impact of usage frequency on omission and regularization. Still, the limited size of ICE and difficulties of handling features that cannot be directly searched for in GloWbE needed to be accounted for.

Other corpora of spoken language for the varieties of interest do exist. Let us briefly elaborate on the reasons why they were not used for the analyses conducted in this book. The NIECSSE (Deterding & Low 2001) mentioned in section 5.1 comprises conversations on a one-to-one basis between speakers of English from Singapore and their British English speaking lecturer (Gut 2009b: 265). Collected in the course

of the project *Towards a Reference Grammar of Singapore English*, the GSSEC, in contrast, consists of eight hours and more than 60,000 words of "naturally-occurring spontaneous discourse of native Singapore English speakers, varying along a number of demographic variables, such as age, gender, ethnic group, and education level" (L. Lim 2009). The language data recorded for the GSSEC were used in compiling ICE-SIN, and speaker background information as well as the original recordings are available. With 60,000 words, the GSSEC alone is too small to approach omission or regularization from a usage-based perspective.

The *Hong Kong Corpus of Spoken English* (HKCSE; W. Cheng et al. 2008), a potential alternative for ICE-HK, is hosted by the Research Centre for Professional Communication in English based at the Hong Kong Polytechnic University and currently comprises about 900,000 words (Hong Kong Polytechnic University 2018). It consists of an academic sub-corpus (e.g., lectures, student presentations, and Q&A), a business sub-corpus (e.g., meetings, presentations), a conversation sub-corpus comprising conversations in different social settings, and a public sub-corpus (e.g., interviews, press briefings). The orthographic version of the corpus can be directly searched via the corpus website, the prosodic version comes as a CD-ROM with W. Cheng et al. (2008). Since, to the author's knowledge, no suitable alternative to ICE exists for IndE and neither the NIECSSE nor the GSSEC proved to be valid choices for SgE, it was decided to stick with ICE-HK as well for reasons of comparability. The language data in ICE and GloWbE are more easily compared than if different corpora for the individual varieties had been accounted for. The HKCSE might be a valid stand-alone alternative though.

The web-based experiment proved to be of considerable value as it made it possible to test the perception of omission of verbal past tense and nominal plural marking. To stick with a feasible experimental design, comparisons of regular with irregular verbs or of verbs and nouns ending in a consonant versus a vowel were not accounted for but would be worth incorporating in future studies. Obviously, experiments need to be very carefully designed and analyzed and they are by far not the key to all research questions. However, sensibly used and properly designed, experiments can be a valuable means to investigate contact phenomena.

Contribution to usage-based linguistics and World Englishes research

In combining a traditional approach to World Englishes research with a usage-based account of omission and regularization, this book contributes to both usage-based linguistics and World Englishes research and has implications for both fields. So

far, usage-based accounts have largely focused on traditional L1 varieties of English like BrE and AmE (for examples see section 2.2) and extending the picture to further variety types clearly adds to the usage-based paradigm. From the observation that frequency effects in the contact varieties investigated are partly constrained by substratum transfer and institutionalization (two well-established concepts in World Englishes research), we learn that instead of solely considering usage frequencies in isolation it is worth including other factors that potentially inhibit or promote frequency effects. The fact that HKE is particularly prone to omission of verbal past tense and nominal plural marking in regular verbs and nouns is clear proof of that.

Turning to the contribution of the usage-based reasoning in this book to World Englishes research, the following is worth pointing out: While frequency counts are commonly reported, the impact usage frequencies have on linguistic features and their development has received little attention. Regular and irregular verb and noun paradigms have repeatedly been dealt with in the usage-based literature, so the idea to expand this focus to World Englishes, or in this case to Asian Englishes, was an obvious choice that promised fruitful insights. In fact, the corpus studies revealed that irregular verbs and nouns, which are (or used to be) highly frequent forms, are not often prone to omission and regularization. This is even the case for HKE, which arguably is a learner variety of English. Only a frequency-based comparison of regular and irregular verbs (compare table 5.4 in section 5.4) could show that even when regular and irregular verbs of comparable frequency are focused on, omission clearly prevails in regular verbs. Consonant cluster reduction was ruled out in the analyses. To the author's knowledge, no previous studies have empirically proven this assumption for the varieties considered here.

A plea for intertwined research

This book can be placed in line with a number of recent studies on Asian varieties of English, of which three are mentioned in the following paragraphs. While looking at the same varieties, the studies differ regarding the features discussed and the methodology used (compare also Hansen 2018: 143). They constitute a continuum that ranges from usage-based accounts of language that combine corpus linguistics with web-based experiments (Horch 2017 and this book) to a quantitative approach to null subjects (Schröter 2017) to a historical account of modality in Asian Englishes (Hansen 2018).

Horch (2017) investigates verb-to-noun conversion in the same three varieties considered here by means of corpus analyses (mainly GloWbE because of the low fre-

quency of occurrence of the features) and an experiment comprising a perception and a production task. She shows that substratum transfer and institutionalization complement each other insofar as substratum transfer only occurs when the superstrate provides the morphosyntactic frame for the transfer (ibid.: 251; compare Bao's (2010) usage-based approach to substratum transfer in section 3.1). The more institutionalized a variety is, the stronger constraints from the superstratum hold (resulting in closer proximity to native Englishes such as BrE and AmE) and the less influence the substratum has (ibid.). As regards frequencies of use, Horch (2017: 143, 149, 248) observes a blocking constraint insofar as frequent near-synonymous deverbal nouns block conversion.

Schröter (2017) examines null subjects in HKE, IndE, SgE, and BrE by means of the respective ICE corpora. For SgE, the GSSEC is additionally considered as it contains speaker background information, which ICE-SIN is lacking. The study "joins the growing body of research empirically evaluating predictions made by theoretical approaches" (ibid.: 2) and adopts a comparative approach using multivariate analysis in order to account for structural factors like person, position, and verb type (62). Typological considerations are of key importance in the analyses and reveal that "contact with typologically different languages leaves measurable traces in the grammar of the Asian Englishes [investigated]" (ibid.: 241). Schröter's (2017) findings show that "Singapore English is the most conspicuous maverick structurally, exhibiting numerous constructions that represent direct calques from the local substrates, including their preference for null subjects" (211). Regarding the determinants tested, SgE and HKE show comparatively many null subjects, but the varieties are also characterized by relatively heterogeneous behavior (ibid.). The fact that SgE follows local substrate languages in its preference for null subjects perfectly ties in with the findings obtained in this book: In contrast with HKE, where no such clear pattern is observable, in SgE consonant cluster reduction considerably accounts for omission of verbal past tense marking in consonant environments (see section 5.3).

Hansen (2018), lastly, approaches modality in Asian Englishes from a historical perspective. She uses the ARCHER corpus (ARCHER-3.2 (Lancaster) 2013) as an approximation of BrE as it was spoken at the time when British settlers came to Hong Kong, India, and Singapore, respectively (e.g., ibid.: 162). The ICE corpora, in contrast, serve to investigate current usage patterns of modality in HKE, IndE, and SgE (e.g., ibid.: 91). By comparing uses of "modal and semi-modal verbs of obligation and necessity" (ibid.: 305) in ARCHER and ICE, Hansen (2018) shows

that substrate languages clearly influence the model system in the Asian varieties investigated. The study closes a research gap in using diachronic data to investigate language change. The lack of diachronic corpora was compensated by apparent-time analyses of ARCHER and ICE. Previous studies have tended to describe well-investigated changes in ESL varieties by means of looking at ENL varieties, thereby often completely lacking a quantitative basis. Hansen, in contrast, promotes the use of apparent-time studies for analyses of age-specific variation and language change.

The studies just mentioned provide a valuable glimpse of how contact varieties can be investigated by means of a wide array of methods, ranging from apparent-time analyses to a corpus-based account applying multivariate analysis to mixed methods research combining corpus data with psycholinguistic experiments (Horch 2017 and this book). All the studies consider both substratum transfer and the degree of institutionalization, two well-established factors in World Englishes research, for the same three Asian varieties. While each study is a valid contribution to World Englishes research per se (both content-wise and methodologically), combined the studies invite intertwined research.

Concluding remarks

This book has shown that for its purposes, systematic empirical research was needed that proved as well as disproved theoretical findings in the literature. Feature lists and in-depth qualitative analyses are valuable contributions to the field of World Englishes, but empirical research is of just as much importance. The corpus studies on omission of verbal past tense and nominal plural marking in particular have shown that careful analyses that account for the phonetic environment and for narrative present (omission of verbal past tense marking only) lead to surprisingly low omission rates.

The triangle of substratum transfer, institutionalization, and usage frequency invites further comparative studies to disentangle the impact of those factors on other features commonly accounted for in the World Englishes literature, whether they are simplifying or not. As elaborated on above, usage-based research and World Englishes research can greatly benefit from each other. It remains to be seen how far usage-based research will account for World Englishes (or L2 Englishes, for that matter) in the future and how far World Englishes research incorporates usage-based thinking. As this book has shown, combining both fields is a promising endeavor.

Appendix

A Transcription conventions

Table A.1: Transcription conventions adopted from the ICE Markup Manual for Spoken Texts (Nelson 2002: 12)

markup symbol	meaning
<$A>, <$B>, *etc*	Speaker identification
<I>...</I>	Subtext marker
<#>	Text unit marker
<O>...</O>	Untranscribed text
<?>...<?>	Uncertain transcription
<->...</->	Normative deletion
<+>...</+>	Normative insertion
<=>...</=>	Original normalization
<.>...</.>	Incomplete word
<}>...</}>	Normative replacement
<[>...</[>	Overlapping string
<{>...</{>	Overlapping string set
<,>	Short pause
<„>	Long pause
<(>...</(>	Discontinuous word
<)>...</)>	Normalized disc. word
<X>...</X>	Extra-corpus text
<&>...</&>	Editorial comment
<@>...</@>	Changed name or word
<w>...</w>	Orthographic word
<quote>...</quote>	Quotation
<mention>...</mention>	Mention
<foreign>...</foreign>	Foreign word(s)
<indig>...</indig>	Indigenous word(s)
<unclear>...</unclear>	Unclear word(s)

B Omission of inflectional past tense marking

Table B.1: Number of verbs marked and not marked for past tense in ICE-SIN, ICE-HK, and ICE-IND, by usage type and morphological process

morphological process	ICE-SIN		ICE-HK		ICE-IND	
	marked	not marked	marked	not marked	marked	not marked
formal uses:						
[ɪd] affixation	262	7	193	13	382	1
[t,d] affixation (C-final, + following C)	368	42	366	70	391	14
[t,d] affixation (C-final, + following V)	304	3	318	26	258	0
[t,d] affixation (V-final, + following C)	355	43	320	62	580	15
[t,d] affixation (V-final, + following V)	226	2	182	22	338	0
suppletion	606	2	399	10	488	1
vowel change	1,402	8	1,250	60	2,061	8
vowel change plus [t,d]	768	1	598	13	601	0
functional uses:						
[ɪd] affixation	144	5	106	9	280	1
[t,d] affixation (C-final, + following C)	126	34	112	25	128	7
[t,d] affixation (C-final, + following V)	113	3	87	12	92	0
[t,d] affixation (V-final, + following C)	128	27	112	34	206	10
[t,d] affixation (V-final, + following V)	97	2	63	18	126	0
suppletion	435	0	291	0	273	1
vowel change	808	4	656	41	852	5
vowel change plus [t,d]	605	1	434	2	379	0

Table B.2: Verb sample in ICE (lemmata in alphabetical order)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-SIN	agree	56	1	1.75	17	1	5.56	170
ICE-SIN	allow	61	2	3.17	7	2	22.22	198
ICE-SIN	apply	39	0	0.00	21	0	0.00	133
ICE-SIN	argue	11	0	0.00	0	0	0.00	32
ICE-SIN	begin	34	1	2.86	27	1	3.57	124
ICE-SIN	break	43	0	0.00	19	0	0.00	74
ICE-SIN	call	277	13	4.48	67	10	12.99	660
ICE-SIN	carry	31	0	0.00	4	0	0.00	138
ICE-SIN	catch	35	0	0.00	15	0	0.00	88
ICE-SIN	claim	18	1	5.26	13	1	7.14	53
ICE-SIN	come	354	0	0.00	273	0	0.00	1,440
ICE-SIN	continue	30	0	0.00	12	0	0.00	148
ICE-SIN	deny	9	0	0.00	2	0	0.00	38
ICE-SIN	destroy	6	0	0.00	0	0	0.00	10
ICE-SIN	develop	31	4	11.43	2	2	50.00	133
ICE-SIN	die	37	0	0.00	29	0	0.00	79
ICE-SIN	drive	28	0	0.00	8	0	0.00	75
ICE-SIN	employ	17	0	0.00	0	0	0.00	35
ICE-SIN	enjoy	28	0	0.00	21	0	0.00	115
ICE-SIN	establish	44	4	8.33	3	0	0.00	68
ICE-SIN	expect	94	1	1.05	8	0	0.00	198
ICE-SIN	fall	21	0	0.00	21	0	0.00	95
ICE-SIN	fight	9	1	10.00	5	1	16.67	54
ICE-SIN	finish	59	8	11.94	21	8	27.59	132

(Continued)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-SIN	follow	42	0	0.00	18	0	0.00	148
ICE-SIN	forget	37	0	0.00	29	0	0.00	90
ICE-SIN	go	606	2	0.33	435	0	0.00	3,290
ICE-SIN	grow	31	1	3.13	17	1	5.56	111
ICE-SIN	happen	83	24	22.43	60	15	20.00	342
ICE-SIN	help	36	1	2.70	14	1	6.67	380
ICE-SIN	identify	20	0	0.00	1	0	0.00	47
ICE-SIN	issue	15	0	0.00	0	0	0.00	24
ICE-SIN	join	15	10	40.00	6	9	60.00	138
ICE-SIN	know	137	0	0.00	65	0	0.00	4,601
ICE-SIN	like	13	0	0.00	10	0	0.00	809
ICE-SIN	live	14	0	0.00	7	0	0.00	151
ICE-SIN	maintain	21	1	4.55	3	0	0.00	63
ICE-SIN	mean	47	0	0.00	23	0	0.00	1,638
ICE-SIN	need	42	2	4.55	12	1	7.69	645
ICE-SIN	play	62	0	0.00	21	0	0.00	291
ICE-SIN	pull	15	1	6.25	5	1	16.67	56
ICE-SIN	rely	5	0	0.00	2	0	0.00	34
ICE-SIN	reply	6	0	0.00	3	0	0.00	16
ICE-SIN	see	372	1	0.27	186	0	0.00	2,015
ICE-SIN	seek	14	0	0.00	3	0	0.00	54
ICE-SIN	show	80	0	0.00	41	0	0.00	382
ICE-SIN	sign	35	4	10.26	8	2	20.00	66
ICE-SIN	stay	14	2	12.50	11	2	15.38	173
ICE-SIN	stick	21	1	4.55	2	0	0.00	37
ICE-SIN	stop	35	9	20.45	20	8	28.57	134

(Continued)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-SIN	take	297	3	1.00	149	1	0.67	1,135
ICE-SIN	tell	315	0	0.00	243	0	0.00	867
ICE-SIN	think	357	1	0.28	321	1	0.31	3,017
ICE-SIN	throw	12	1	7.69	6	1	14.29	68
ICE-SIN	view	0	0	0.00	0	0	0.00	7
ICE-SIN	visit	11	0	0.00	10	0	0.00	66
ICE-SIN	want	115	4	3.36	114	4	3.39	1,507
ICE-SIN	watch	27	5	15.63	15	4	21.05	204
ICE-SIN	wear	6	0	0.00	2	0	0.00	91
ICE-SIN	wish	0	0	0.00	0	0	0.00	58
ICE-HK	agree	59	3	4.84	19	3	13.64	289
ICE-HK	allow	52	6	10.34	3	1	25.00	200
ICE-HK	apply	37	6	13.95	3	2	40.00	179
ICE-HK	argue	11	1	8.33	8	0	0.00	66
ICE-HK	begin	36	3	7.69	27	2	6.90	127
ICE-HK	break	48	0	0.00	18	0	0.00	102
ICE-HK	call	253	50	16.50	33	8	19.51	644
ICE-HK	carry	29	3	9.38	10	2	16.67	145
ICE-HK	catch	23	0	0.00	5	0	0.00	67
ICE-HK	claim	17	1	5.56	8	1	11.11	63
ICE-HK	come	248	19	7.12	183	19	9.41	1,638
ICE-HK	continue	18	0	0.00	4	0	0.00	216
ICE-HK	deny	12	1	7.69	5	0	0.00	24
ICE-HK	destroy	5	1	16.67	2	1	33.33	11
ICE-HK	develop	49	7	12.50	5	2	28.57	148
ICE-HK	die	22	3	12.00	14	3	17.65	84

(Continued)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-HK	drive	29	0	0.00	13	0	0.00	122
ICE-HK	employ	10	0	0.00	1	0	0.00	35
ICE-HK	enjoy	7	3	30.00	4	2	33.33	163
ICE-HK	establish	39	2	4.88	1	0	0.00	79
ICE-HK	expect	53	0	0.00	8	0	0.00	185
ICE-HK	fall	29	1	3.33	18	1	5.26	77
ICE-HK	fight	9	1	10.00	5	1	16.67	72
ICE-HK	finish	58	10	14.71	20	6	23.08	209
ICE-HK	follow	121	4	3.20	6	0	0.00	261
ICE-HK	forget	38	6	13.64	26	6	18.75	125
ICE-HK	go	399	10	2.44	291	0	0.00	4,283
ICE-HK	grow	21	0	0.00	7	0	0.00	100
ICE-HK	happen	93	21	18.42	76	17	18.28	424
ICE-HK	help	15	7	31.82	6	5	45.45	323
ICE-HK	identify	22	1	4.35	10	1	9.09	40
ICE-HK	issue	21	0	0.00	4	0	0.00	40
ICE-HK	join	35	11	23.91	13	5	27.78	222
ICE-HK	know	103	2	1.90	43	0	0.00	5,438
ICE-HK	like	8	0	0.00	5	0	0.00	1,780
ICE-HK	live	19	9	32.14	12	8	40.00	471
ICE-HK	maintain	19	0	0.00	3	0	0.00	79
ICE-HK	mean	27	3	10.00	13	0	0.00	2,314
ICE-HK	need	33	0	0.00	13	0	0.00	763
ICE-HK	play	12	5	29.41	5	4	44.44	329
ICE-HK	pull	14	1	6.67	8	0	0.00	52
ICE-HK	rely	4	0	0.00	2	0	0.00	40

(Continued)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-HK	reply	2	0	0.00	2	0	0.00	9
ICE-HK	see	308	4	1.28	129	0	0.00	2,369
ICE-HK	seek	8	0	0.00	1	0	0.00	68
ICE-HK	show	77	3	3.75	20	3	13.04	409
ICE-HK	sign	23	4	14.81	13	0	0.00	81
ICE-HK	stay	29	8	21.62	27	8	22.86	417
ICE-HK	stick	8	1	11.11	3	0	0.00	39
ICE-HK	stop	18	5	21.74	13	4	23.53	149
ICE-HK	take	345	22	5.99	171	12	6.56	1,694
ICE-HK	tell	324	5	1.52	237	2	0.84	1,078
ICE-HK	think	216	5	2.26	178	0	0.00	5,463
ICE-HK	throw	14	1	6.67	7	0	0.00	70
ICE-HK	view	5	0	0.00	1	0	0.00	12
ICE-HK	visit	34	6	15.00	13	2	13.33	172
ICE-HK	want	73	7	8.75	72	7	8.86	1,818
ICE-HK	watch	23	4	14.81	5	3	37.50	229
ICE-HK	wear	14	0	0.00	6	0	0.00	83
ICE-HK	wish	3	0	0.00	3	0	0.00	77
ICE-IND	agree	38	0	0.00	9	0	0.00	123
ICE-IND	allow	53	0	0.00	9	0	0.00	120
ICE-IND	apply	37	0	0.00	8	0	0.00	86
ICE-IND	argue	7	0	0.00	3	0	0.00	35
ICE-IND	begin	58	0	0.00	37	0	0.00	185
ICE-IND	break	55	0	0.00	8	0	0.00	80
ICE-IND	call	405	5	1.22	58	1	1.69	683
ICE-IND	carry	27	0	0.00	6	0	0.00	134

(Continued)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-IND	catch	36	0	0.00	13	0	0.00	57
ICE-IND	claim	14	0	0.00	5	0	0.00	43
ICE-IND	come	605	0	0.00	381	0	0.00	1,982
ICE-IND	continue	27	0	0.00	15	0	0.00	167
ICE-IND	deny	12	0	0.00	1	0	0.00	23
ICE-IND	destroy	21	0	0.00	4	0	0.00	39
ICE-IND	develop	81	1	1.22	18	0	0.00	127
ICE-IND	die	34	0	0.00	20	0	0.00	105
ICE-IND	drive	9	0	0.00	2	0	0.00	25
ICE-IND	employ	23	0	0.00	3	0	0.00	37
ICE-IND	enjoy	19	0	0.00	16	0	0.00	123
ICE-IND	establish	47	0	0.00	8	0	0.00	72
ICE-IND	expect	53	0	0.00	9	0	0.00	116
ICE-IND	fall	46	0	0.00	31	0	0.00	105
ICE-IND	fight	14	0	0.00	7	0	0.00	64
ICE-IND	finish	41	1	2.38	19	1	5.00	86
ICE-IND	follow	45	0	0.00	14	0	0.00	208
ICE-IND	forget	22	5	18.52	13	5	27.78	63
ICE-IND	go	488	1	0.20	273	1	0.36	2,708
ICE-IND	grow	15	0	0.00	3	0	0.00	51
ICE-IND	happen	171	14	7.57	126	12	8.70	456
ICE-IND	help	20	0	0.00	12	0	0.00	203
ICE-IND	identify	45	0	0.00	12	0	0.00	94
ICE-IND	issue	43	0	0.00	1	0	0.00	62
ICE-IND	join	54	2	3.57	35	1	2.78	138
ICE-IND	know	248	0	0.00	36	0	0.00	2,634

(Continued)

corpus	lemma	formal uses			functional uses			lemma token freq. (ICE)
		marked	not marked	omission rate %	marked	not marked	omission rate %	
ICE-IND	like	19	0	0.00	16	0	0.00	148
ICE-IND	live	15	0	0.00	11	0	0.00	148
ICE-IND	maintain	29	0	0.00	2	0	0.00	85
ICE-IND	mean	36	0	0.00	10	0	0.00	1,152
ICE-IND	need	60	0	0.00	21	0	0.00	315
ICE-IND	play	104	0	0.00	48	0	0.00	275
ICE-IND	pull	10	0	0.00	6	0	0.00	29
ICE-IND	rely	2	0	0.00	1	0	0.00	9
ICE-IND	reply	9	0	0.00	7	0	0.00	20
ICE-IND	see	374	1	0.27	144	0	0.00	1,941
ICE-IND	seek	16	0	0.00	6	0	0.00	51
ICE-IND	show	89	0	0.00	20	0	0.00	279
ICE-IND	sign	31	0	0.00	3	0	0.00	45
ICE-IND	stay	27	0	0.00	21	0	0.00	233
ICE-IND	stick	13	1	7.14	3	0	0.00	24
ICE-IND	stop	19	5	20.83	9	1	10.00	74
ICE-IND	take	571	1	0.17	170	0	0.00	1,767
ICE-IND	tell	313	0	0.00	211	0	0.00	839
ICE-IND	think	200	0	0.00	139	0	0.00	1,832
ICE-IND	throw	21	0	0.00	12	0	0.00	60
ICE-IND	view	2	0	0.00	0	0	0.00	7
ICE-IND	visit	58	1	1.69	45	1	2.17	154
ICE-IND	want	211	0	0.00	205	0	0.00	944
ICE-IND	watch	10	1	9.09	4	1	20.00	134
ICE-IND	wear	10	0	0.00	5	0	0.00	53
ICE-IND	wish	5	0	0.00	2	0	0.00	56

Table B.3: Verb sample in GloWbE (lemmata in alphabetical order)

corpus	lemma	hits for LEMMA.[v*]	not marked (sample)	not marked (corpus)	marked (.[vvd])	omission rate
SG	agree	7,302	1	36.51	1,116	3.17
SG	allow	5,949	0	0.00	914	0.00
SG	apply	5,599	0	0.00	418	0.00
SG	argue	1,411	0	0.00	383	0.00
SG	begin	3,294	0	0.00	3,765	0.00
SG	break	2,843	0	0.00	1,446	0.00
SG	call	6,801	2	68.01	1,971	3.34
SG	carry	3,461	0	0.00	557	0.00
SG	catch	2,883	0	0.00	997	0.00
SG	claim	1,838	0	0.00	835	0.00
SG	come	25,051	0	0.00	13,037	0.00
SG	continue	8,990	0	0.00	1,855	0.00
SG	deny	732	0	0.00	204	0.00
SG	destroy	772	0	0.00	204	0.00
SG	develop	4,473	0	0.00	653	0.00
SG	die	3,320	0	0.00	2,046	0.00
SG	drive	3,007	0	0.00	767	0.00
SG	employ	582	1	2.91	187	1.53
SG	enjoy	8,844	0	0.00	2,458	0.00
SG	establish	1,370	0	0.00	544	0.00
SG	expect	5,688	1	28.44	913	3.02
SG	fall	3,511	0	0.00	2,563	0.00
SG	fight	2,434	0	0.00	385	0.00
SG	finish	2,136	3	32.04	1,147	2.72
SG	follow	6,081	0	0.00	1,175	0.00
SG	forget	3,978	4	79.56	1,088	6.81
SG	go	43,204	0	0.00	13,485	0.00
SG	grow	4,667	0	0.00	1,949	0.00
SG	happen	6,057	5	151.43	3,960	3.68

(Continued)

corpus	lemma	hits for LEMMA.[v*]	not marked (sample)	not marked (corpus)	marked (.[vvd])	omission rate
SG	help	21,767	1	108.84	1,755	5.84
SG	identify	1,794	0	0.00	95	0.00
SG	issue	705	0	0.00	373	0.00
SG	join	3,786	0	0.00	1,649	0.00
SG	know	49,995	0	0.00	5,423	0.00
SG	like	25,804	0	0.00	2,175	0.00
SG	live	9,277	0	0.00	1,210	0.00
SG	maintain	2,811	0	0.00	209	0.00
SG	mean	9,353	0	0.00	1,831	0.00
SG	need	33,880	0	0.00	2,710	0.00
SG	play	8,306	1	41.53	2,130	1.91
SG	pull	1,916	1	9.58	673	1.40
SG	rely	1,251	0	0.00	87	0.00
SG	reply	670	0	0.00	1,142	0.00
SG	see	47,947	0	0.00	8,331	0.00
SG	seek	2,784	0	0.00	437	0.00
SG	show	9,133	0	0.00	2,864	0.00
SG	sign	1,752	0	0.00	637	0.00
SG	stay	9,034	1	45.17	1,616	2.72
SG	stick	1,709	0	0.00	324	0.00
SG	stop	7,271	2	72.71	1341	5.14
SG	take	42,331	0	0.00	11,791	0.00
SG	tell	12,192	0	0.00	8,234	0.00
SG	think	48,838	0	0.00	8,529	0.00
SG	throw	1,809	0	0.00	643	0.00
SG	view	1,791	0	0.00	128	0.00
SG	visit	6,428	0	0.00	1,352	0.00
SG	want	41,294	0	0.00	8,131	0.00
SG	watch	7,772	1	38.86	1,987	1.92
SG	wear	3,306	0	0.00	685	0.00
SG	wish	5,913	0	0.00	575	0.00
HK	agree	4,204	0	0.00	1,232	0.00

(Continued)

corpus	lemma	hits for LEMMA.[v*]	not marked (sample)	not marked (corpus)	marked (.[vvd])	omission rate
HK	allow	5,849	1	29.25	807	3.50
HK	apply	6,804	0	0.00	403	0.00
HK	argue	1,077	0	0.00	460	0.00
HK	begin	3,610	1	18.05	6,059	0.30
HK	break	2,261	0	0.00	1,035	0.00
HK	call	5,635	1	28.18	1,970	1.41
HK	carry	3,851	0	0.00	764	0.00
HK	catch	1,685	0	0.00	673	0.00
HK	claim	1,592	1	7.96	832	0.95
HK	come	19,848	0	0.00	11,854	0.00
HK	continue	8,314	0	0.00	2,141	0.00
HK	deny	610	0	0.00	279	0.00
HK	destroy	781	0	0.00	260	0.00
HK	develop	5,834	0	0.00	926	0.00
HK	die	1,895	0	0.00	2,443	0.00
HK	drive	2,286	0	0.00	652	0.00
HK	employ	746	0	0.00	270	0.00
HK	enjoy	6,743	0	0.00	1,511	0.00
HK	establish	2,474	0	0.00	1,083	0.00
HK	expect	4,138	0	0.00	789	0.00
HK	fall	2,615	0	0.00	2,368	0.00
HK	fight	1,859	0	0.00	295	0.00
HK	finish	1,817	11	99.94	1,035	8.81
HK	follow	5,407	0	0.00	1,239	0.00
HK	forget	2,564	1	12.82	546	2.29
HK	go	29,844	0	0.00	11,556	0.00
HK	grow	3,775	0	0.00	2,070	0.00
HK	happen	4,122	3	61.83	2,610	2.31
HK	help	19,736	0	0.00	1,741	0.00
HK	identify	2,478	0	0.00	125	0.00
HK	issue	1,119	0	0.00	938	0.00
HK	join	3,815	2	38.15	2,243	1.67

(Continued)

corpus	lemma	hits for LEMMA.[v*]	not marked (sample)	not marked (corpus)	marked (.[vvd])	omission rate
HK	know	31,314	0	0.00	4,532	0.00
HK	like	15,486	0	0.00	1,269	0.00
HK	live	8,168	2	81.68	1,725	4.52
HK	maintain	3,895	0	0.00	259	0.00
HK	mean	5,880	0	0.00	1,447	0.00
HK	need	28,344	0	0.00	2,491	0.00
HK	play	7,101	0	0.00	1,666	0.00
HK	pull	1,411	2	14.11	618	2.23
HK	rely	1,458	0	0.00	75	0.00
HK	reply	523	0	0.00	1,028	0.00
HK	see	35,270	0	0.00	6,325	0.00
HK	seek	2,805	0	0.00	534	0.00
HK	show	8,050	0	0.00	3,070	0.00
HK	sign	1,624	0	0.00	835	0.00
HK	stay	6,099	0	0.00	1,007	0.00
HK	stick	1,262	0	0.00	207	0.00
HK	stop	5,120	1	25.60	1,101	2.27
HK	take	33,712	0	0.00	10,457	0.00
HK	tell	8,874	0	0.00	6,477	0.00
HK	think	29,290	0	0.00	6,172	0.00
HK	throw	1,075	0	0.00	509	0.00
HK	view	1,728	0	0.00	160	0.00
HK	visit	5,660	0	0.00	1,620	0.00
HK	want	28,175	0	0.00	6,027	0.00
HK	watch	3,325	1	16.63	821	1.98
HK	wear	2,773	0	0.00	635	0.00
HK	wish	4,469	0	0.00	434	0.00
IN	agree	11,993	0	0.00	2,363	0.00
IN	allow	13,547	0	0.00	2,073	0.00
IN	apply	9,869	0	0.00	782	0.00
IN	argue	2,604	0	0.00	1,024	0.00

(Continued)

corpus	lemma	hits for LEMMA.[v*]	not marked (sample)	not marked (corpus)	marked (.[vvd])	omission rate
IN	begin	8,561	0	0.00	11,198	0.00
IN	break	5,969	0	0.00	2,910	0.00
IN	call	14,668	0	0.00	4,827	0.00
IN	carry	8,598	0	0.00	1,908	0.00
IN	catch	3,568	0	0.00	1,580	0.00
IN	claim	5,129	0	0.00	3,324	0.00
IN	come	61,454	0	0.00	31,474	0.00
IN	continue	16,856	0	0.00	4,357	0.00
IN	deny	2,104	0	0.00	1,086	0.00
IN	destroy	2,789	0	0.00	808	0.00
IN	develop	9,835	0	0.00	1,696	0.00
IN	die	5,760	0	0.00	5,604	0.00
IN	drive	4,947	0	0.00	1,246	0.00
IN	employ	1,251	0	0.00	440	0.00
IN	enjoy	12,241	0	0.00	2,517	0.00
IN	establish	3,795	0	0.00	1,554	0.00
IN	expect	10,773	0	0.00	1,842	0.00
IN	fall	6,723	0	0.00	4,918	0.00
IN	fight	6,154	0	0.00	1,278	0.00
IN	finish	2,836	3	21.27	1,674	1.25
IN	follow	15,096	0	0.00	3,372	0.00
IN	forget	8,379	1	20.95	1,408	1.47
IN	go	75,239	0	0.00	24,180	0.00
IN	grow	8,727	0	0.00	3,755	0.00
IN	happen	14,016	3	105.12	8,208	1.26
IN	help	46,712	0	0.00	4,122	0.00
IN	identify	5,265	0	0.00	393	0.00
IN	issue	2,014	0	0.00	1,394	0.00
IN	join	7,831	1	19.58	3,908	0.50
IN	know	88,402	0	0.00	9,699	0.00
IN	like	37,310	0	0.00	2,894	0.00
IN	live	19,289	0	0.00	3,669	0.00

(Continued)

corpus	lemma	hits for LEMMA.[v*]	not marked (sample)	not marked (corpus)	marked (.[vvd])	omission rate
IN	maintain	7,309	0	0.00	903	0.00
IN	mean	15,862	0	0.00	3,752	0.00
IN	need	72,804	0	0.00	5,328	0.00
IN	play	21,528	0	0.00	6,360	0.00
IN	pull	3,406	1	8.52	1,317	0.64
IN	rely	2,341	0	0.00	189	0.00
IN	reply	2,473	0	0.00	2,917	0.00
IN	see	83,946	0	0.00	14,302	0.00
IN	seek	5,985	0	0.00	1,890	0.00
IN	show	18,578	0	0.00	6,405	0.00
IN	sign	3,049	1	7.62	1,315	0.58
IN	stay	13,861	0	0.00	2,083	0.00
IN	stick	2,864	0	0.00	426	0.00
IN	stop	14,168	0	0.00	2,349	0.00
IN	take	89,312	0	0.00	24,774	0.00
IN	tell	22,592	0	0.00	19,934	0.00
IN	think	72,869	0	0.00	11,626	0.00
IN	throw	3,532	0	0.00	1,362	0.00
IN	view	3,555	0	0.00	335	0.00
IN	visit	10,733	0	0.00	2,997	0.00
IN	want	78,063	0	0.00	14,857	0.00
IN	watch	10,531	2	52.66	1,726	2.96
IN	wear	5,148	0	0.00	1,110	0.00
IN	wish	10,225	0	0.00	997	0.00

C Omission of inflectional noun plural marking

Table C.1: Noun sample in ICE (lemmata in alphabetical order)

corpus	lemma	marked	not marked	omission rate %	lemma token freq. (ICE)
ICE-SIN	boy	71	0	0.00	176
ICE-SIN	day	267	2	0.74	701
ICE-SIN	detail	48	2	4.00	64
ICE-SIN	eye	38	0	0.00	94
ICE-SIN	friend	111	1	0.89	320
ICE-SIN	girl	46	0	0.00	130
ICE-SIN	hour	107	1	0.93	207
ICE-SIN	parent	174	1	0.57	204
ICE-SIN	point	104	1	0.95	739
ICE-SIN	problem	203	3	1.46	612
ICE-SIN	question	114	1	0.87	474
ICE-SIN	reason	62	3	4.62	205
ICE-SIN	school	136	0	0.00	481
ICE-SIN	shoe	24	2	7.69	31
ICE-SIN	student	332	10	2.92	462
ICE-SIN	teacher	108	1	0.92	220
ICE-SIN	term	207	3	1.43	327
ICE-SIN	thing	652	18	2.69	1,355
ICE-SIN	way	89	2	2.20	794
ICE-SIN	year	625	3	0.48	1,346
ICE-HK	boy	29	0	0.00	137
ICE-HK	day	190	5	2.56	576
ICE-HK	detail	68	1	1.45	101

(Continued)

corpus	lemma	marked	not marked	omission rate %	lemma token freq. (ICE)
ICE-HK	eye	28	3	9.68	86
ICE-HK	friend	209	17	7.52	365
ICE-HK	girl	86	6	6.52	278
ICE-HK	hour	134	8	5.63	296
ICE-HK	parent	107	23	17.69	136
ICE-HK	point	70	5	6.67	694
ICE-HK	problem	204	16	7.27	665
ICE-HK	question	156	20	11.36	651
ICE-HK	reason	40	5	11.11	287
ICE-HK	school	97	22	18.49	976
ICE-HK	shoe	11	0	0.00	13
ICE-HK	student	376	69	15.51	704
ICE-HK	teacher	75	4	5.06	385
ICE-HK	term	189	5	2.58	398
ICE-HK	thing	442	30	6.36	924
ICE-HK	way	54	6	10.00	671
ICE-HK	year	607	15	2.41	1,498
ICE-IND	boy	72	2	2.70	149
ICE-IND	day	343	0	0.00	959
ICE-IND	detail	50	0	0.00	73
ICE-IND	eye	31	0	0.00	65
ICE-IND	friend	129	3	2.27	256
ICE-IND	girl	119	1	0.83	209
ICE-IND	hour	141	0	0.00	232
ICE-IND	parent	115	1	0.86	123
ICE-IND	point	107	4	3.60	709
ICE-IND	problem	192	7	3.52	582
ICE-IND	question	86	5	5.49	465
ICE-IND	reason	48	5	9.43	203
ICE-IND	school	92	0	0.00	490
ICE-IND	shoe	14	0	0.00	14

(Continued)

corpus	lemma	marked	not marked	omission rate %	lemma token freq. (ICE)
ICE-IND	student	415	16	3.71	547
ICE-IND	teacher	152	9	5.59	412
ICE-IND	term	154	4	2.53	239
ICE-IND	thing	557	10	1.76	1,214
ICE-IND	way	72	6	7.69	773
ICE-IND	year	499	6	1.19	1,032

Table C.2: Noun sample in GloWbE (lemmata in alphabetical order)

corpus	lemma	hits for LEMMA.[n*]	not marked (sample)	not marked (corpus)	marked (.[n*])	omission rate
SG	boy	4,375	0	0.00	3,236	0.00
SG	day	38,284	0	0.00	19,955	0.00
SG	detail	1,528	0	0.00	5,012	0.00
SG	eye	5,422	0	0.00	7,818	0.00
SG	friend	8,177	1	40.89	14,074	0.29
SG	girl	8,745	0	0.00	7,327	0.00
SG	hour	4,968	0	0.00	10,783	0.00
SG	parent	1,694	1	8.47	8,402	0.10
SG	point	15,349	0	0.00	5,161	0.00
SG	problem	11,610	0	0.00	8,272	0.00
SG	question	9,550	1	47.75	7,076	0.67
SG	reason	10,238	0	0.00	5,031	0.00
SG	school	17,615	0	0.00	4,408	0.00
SG	shoe	1,008	1	5.04	3,168	0.16
SG	student	4,976	1	24.88	12,584	0.20
SG	teacher	3,520	0	0.00	3,398	0.00
SG	term	5,563	0	0.00	5,803	0.00
SG	thing	19,127	0	0.00	26,892	0.00
SG	way	45,614	0	0.00	7,418	0.00

(Continued)

corpus	lemma	hits for LEMMA.[n*]	not marked (sample)	not marked (corpus)	marked (.[n*])	omission rate
SG	year	35,974	0	0.00	39,564	0.00
HK	boy	2,792	0	0.00	1,948	0.00
HK	day	32,207	0	0.00	17,324	0.00
HK	detail	1,806	2	18.06	5,985	0.30
HK	eye	3,612	0	0.00	5,317	0.00
HK	friend	5,883	0	0.00	10,150	0.00
HK	girl	4,224	0	0.00	3,258	0.00
HK	hour	4,291	1	21.46	9,197	0.23
HK	parent	1,490	1	7.45	6,796	0.11
HK	point	13,227	0	0.00	4,813	0.00
HK	problem	12,104	0	0.00	9,508	0.00
HK	question	8,432	0	0.00	6,656	0.00
HK	reason	9,081	0	0.00	4,976	0.00
HK	school	18,784	0	0.00	5,603	0.00
HK	shoe	986	3	14.79	4,140	0.36
HK	student	6,513	2	65.13	19,739	0.33
HK	teacher	3,953	0	0.00	4,004	0.00
HK	term	5,691	0	0.00	6,171	0.00
HK	thing	13,246	0	0.00	19,665	0.00
HK	way	37,354	0	0.00	6,940	0.00
HK	year	37,302	0	0.00	42,766	0.00
IN	boy	7,794	0	0.00	5,125	0.00
IN	day	78,271	0	0.00	41,225	0.00
IN	detail	3,919	1	9.80	12,755	0.08
IN	eye	9,016	0	0.00	13,449	0.00
IN	friend	14,075	0	0.00	22,783	0.00
IN	girl	14,418	0	0.00	9,570	0.00
IN	hour	7,740	0	0.00	17,404	0.00
IN	parent	3,211	1	8.03	15,766	0.05
IN	point	34,875	0	0.00	12,026	0.00
IN	problem	31,990	1	79.98	22,158	0.36

(Continued)

corpus	lemma	hits for LEMMA.[n*]	not marked (sample)	not marked (corpus)	marked (.[n*])	omission rate
IN	question	26,889	1	67.22	16,833	0.40
IN	reason	24,604	1	61.51	13,091	0.47
IN	school	25,965	0	0.00	10,089	0.00
IN	shoe	1,048	1	2.62	5,082	0.05
IN	student	12,325	2	61.63	26,386	0.23
IN	teacher	6,751	0	0.00	6,804	0.00
IN	term	14,578	0	0.00	9,876	0.00
IN	thing	37,235	1	93.09	56,182	0.17
IN	way	98,437	0	0.00	18,159	0.00
IN	year	69,952	2	349.76	90,722	0.38

D Regularization

Table D.1: Uncountable nouns used as countable nouns in GloWbE: Number of regularized plural and singular forms (five occurrences or more in any corpus, in alphabetical order)

lemma	GloWbE SG		GloWbE HK		GloWbE IN		GloWbE GB		GloWbE US	
	pluralized	singular	pluralized	singular	pluralized	singular	pluralized	singular	pluralized	singular
advertising	1	2,261	5	2,630	1	4,498	5	17,646	2	15,290
advice	107	4,590	117	4,440	163	7,538	167	54,918	145	34,407
anger	9	1,016	4	780	6	3,372	20	10,113	22	10,912
assistance	1	1,642	8	2,400	6	3,675	14	10,861	8	16,865
blood	15	4,890	7	4,055	19	9,793	236	39,785	85	40,132
bread	200	2,582	96	1,463	199	1,950	431	9,941	598	11,733
cash	1	4,134	2	3,715	5	7,471	7	29,472	4	24,197
consciousness	4	684	19	1,638	15	7,680	52	7,940	57	10,527
coordination	0	429	2	614	0	1,238	1	2,037	6	2,850
corruption	13	1,000	15	1,236	107	12,005	88	9,810	224	9,825
coverage	6	1,486	6	2,342	23	4,174	27	16,217	225	24,522
data	2	9,303	6	13,618	30	34,322	13	104,583	14	116,944
education	15	8,712	28	12,915	42	26,428	158	69,072	461	80,149
equipment	140	2,624	261	4,531	541	5,516	202	19,958	85	16,689
fun	2	4,340	14	2,689	4	5,212	18	26,214	12	26,745
furniture	25	987	14	1,879	10	1,908	18	7,355	15	5,785
happiness	0	2,233	0	1,728	1	5,825	7	9,990	9	13,963
health	1	10,820	0	11,296	7	23,180	11	101,820	3	129,991
homework	4	558	13	649	0	1,149	17	3,490	3	5,634
immigration	19	1,499	7	2,904	7	1,939	25	13,340	31	15,570

(Continued)

lemma	GloWbE SG		GloWbE HK		GloWbE IN		GloWbE GB		GloWbE US	
	pluralized	singular	pluralized	singular	pluralized	singular	pluralized	singular	pluralized	singular
importance	1	2,516	0	3,159	4	9,430	4	25,584	5	21,372
inflation	4	1,381	4	1,432	7	4,835	16	12,268	15	12,711
information	27	19,082	45	26,086	111	47,395	151	154,003	135	163,820
jewellery	4	334	0	934	15	1,839	4	3,479	1	307
jewelry	9	667	18	1,544	56	1,663	46	1,377	16	3,185
knowledge	5	6,101	29	7,996	18	22,799	62	57,797	58	55,168
learning	30	3,887	17	4,506	142	6,790	175	27,471	139	22,004
legislation	33	845	66	1,660	215	2,505	111	17,578	75	17,149
luck	6	2,912	2	1,577	6	3,980	15	22,825	10	20,379
machinery	8	405	20	1,051	39	1,711	24	3,415	28	2,701
marketing	0	7,298	0	6,870	0	12,990	2	35,747	7	31,042
milk	3	3,308	3	2,110	10	5,021	66	12,557	77	12,399
music	12	10,354	3	11,170	5	19,566	89	103,205	58	77,772
planning	3	2,062	0	2,978	8	5,522	3	18,382	3	13,261
pollution	4	670	8	2,317	30	2,545	17	4,789	37	5,939
privacy	3	1,706	0	2,224	1	2,915	4	12,148	14	14,367
recognition	11	1,373	17	2,226	45	3,440	46	12,340	59	10,823
rice	7	5,103	7	2,458	6	4,478	13	6,344	29	12,997
software	21	4,232	42	5,057	293	21,054	83	36,498	127	40,076
steel	7	967	5	2,223	38	3,599	83	8,321	68	7,270
storage	8	2,016	10	2,403	77	5,421	11	11,850	16	13,276
stuff	294	5,172	82	2,846	265	6,592	163	55,271	284	66,040
thinking	2	2,612	18	2,601	6	6,028	10	26,004	14	28,001
traffic	2	3,491	8	3,944	22	9,548	9	24,254	7	23,449

(Continued)

lemma	GloWbE SG		GloWbE HK		GloWbE IN		GloWbE GB		GloWbE US	
	pluralized	singular	pluralized	singular	pluralized	singular	pluralized	singular	pluralized	singular
training	120	6,464	126	7,738	161	13,013	169	55,636	364	37,844
violence	0	1,192	0	1,470	1	8,719	14	27,007	18	30,512
weather	12	2,829	7	2,970	12	4,456	176	26,430	21	22,145

E Testing the perception of lack of inflectional marking

E.1 Background questions

1. How old are you? [blank]
2. What is your sex? [blank]
3. Which country are you living in now? [blank]
4. How many months have you been abroad in other English-speaking countries (not including holidays)? [options to tick: never, less than 1 month, 1–3 months, 4–6 months, 7–9 months, longer than 9 months]
5. Which countries were that? [blank]
6. How many months have you been abroad overall (not including holidays)? [options to tick: never, less than 1 month, 1–3 months, 4–6 months, 7–9 months, longer than 9 months]
7. What is your native language? [blank]
8. Which other language(s) and/or dialect(s) do you speak? Please provide them in the order of proficiency, starting with the one you feel most proficient in. [blank]
9. At which age did you start learning English? [blank]
10. Which language(s) and/or dialect(s) does your family mainly use at home? [blank]
11. What is your highest completed level of education? [blank]
12. What is your current status? [options to tick: undergraduate student, postgraduate student, other: [blank]]

13. Have you taken classes in linguistics? [options to tick: no, yes]
14. Are you left-handed or right-handed? [options to tick: left-handed, right-handed]

E.2 Follow-up questions

1. Do you have an idea what language feature(s) the tasks were about? [blank]
2. Is that something you are familiar with? [blank]
3. With regard to the judgments you provided in the second task, how confident did you feel overall when making your decisions? [blank]
4. Do you have further comments? [blank]

E.3 Task design

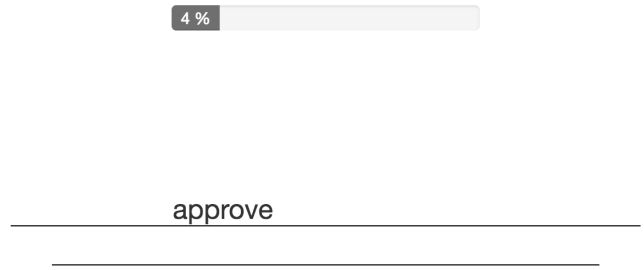

Figure E.1: Design of the self-paced reading task, with a bar on top that shows the participant's progress in the task

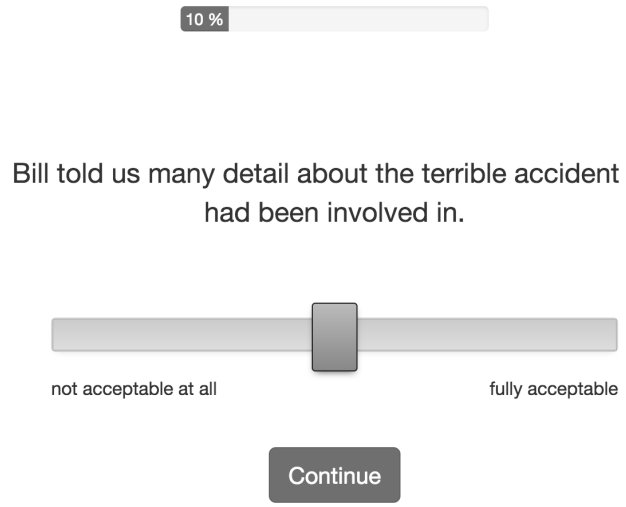

Figure E.2: Design of the acceptability judgment task, with a bar on top that shows the participant's progress in the task; the progress was measured for each task individually

Figure E.3: Design of the comprehension questions in both the self-paced reading and the acceptability judgment task; the answer option "yes" (in grey) was chosen here; the chosen answer option turned green in case the answer was correct and red in case it was wrong

Table E.1: Stimulus lists

list	task	order of conditions (set types V and N)
list 1	SPR	1 · **2** · 3 · **4**
	AJT	**4** · 3 · **2** · 1
list 2	SPR	**2** · **4** · 1 · 3
	AJT	3 · 1 · **4** · **2**
list 3	SPR	**4** · 3 · **2** · 1
	AJT	1 · **2** · 3 · **4**
list 4	SPR	3 · 1 · **4** · **2**
	AJT	**2** · **4** · 1 · 3

Table E.2: Stimuli

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
V	1	1	✓	✓	*Last weekend* Ben and Vanessa supervised eight kids and *played* ice hockey with them on a frozen lake. Q (task 1): Did Ben play ice hockey? (yes)
V	2	1	✓	✓	*Yesterday* they talked about their next steps and unanimously *agreed* on booking a flight to Canada. Q (task 1): Will they book a flight to Canada? (yes)
V	3	1	✓	✓	*Last time* they called us to the counter and *allowed* us to cancel the tickets free of charge. Q (task 2): Did we have to pay a fee to cancel the tickets? (no)
V	4	1	✓		*One week ago* the police departments adopted a proven strategy and *issued* a description of the criminal.
V	5	1	✓	✓	*Yesterday* they rehearsed a presentation and *stayed* in the library half the night. Q (task 2): Did they rehearse a presentation in the library? (yes)
V	6	1	✓	✓	*Last weekend* Joe and Lisa cooked a meal for us and *showed* us their awesome holiday photos. Q (task 1): Did we cook for Joe and Lisa? (no)
V	7	1	✓		*The other day* Joseph and Billy joined a guided tour and *prayed* in vain that they would be able to keep the pace. Q (task 1): Were Joseph and Billy able to keep the pace? (no)
V	8	1	✓	✓	*Back then* we feared we might lose our jobs and *continued* working with even more effort. Q (task 2): Did we fear to lose our jobs? (yes)
V	9	1	✓	✓	*Recently* Jim and Lea rejected an extension of their contracts and successfully *applied* for alternative positions. Q (task 2): Did Jim and Lea reject alternative positions? (no)
V	10	1	✓		*Back then* Elena's dresses often convinced the critics and always *varied* in shape to fit the occasion.
V	11	1	✓	✓	*Last Monday* Marcus surprised Angela with a bouquet of roses and *carried* it inside to put the flowers in a vase.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
					Q (task 1): Did Angela surprise Marcus with roses? (no)
V	12	1	✓	✓	*Recently* Isabel and Lenny moved away and *died* in a car crash close to their new home.
					Q (task 1): Did Isabel die in a car crash? (yes)
V	13	1	✓		*Recently* Jenny and Lilly attended a conference and *tied* it in with a stopover at their first apartment.
					Q (task 1): Did Jenny and Lilly stop at their first apartment? (yes)
V	14	1	✓	✓	*Back then* the two divisions surpassed their competitors and immediately *employed* eleven further trainees.
					Q (task 2): Were the divisions surpassed by their competitors? (no)
V	15	1	✓	✓	*In 2010* we decided against selling our flat but *viewed* a place nearby nevertheless.
					Q (task 2): Did we sell our flat? (no)
V	16	1	✓	✓	*The other day* we discovered a dead bird and *buried* it in our garden under the apple tree.
					Q (task 1): Do we have an apple tree in the garden? (yes)
V	17	1	✓		*A week ago* Ellen and George received a weird email and *replied* instead of simply ignoring it.
					Q (task 1): Did Ellen and George ignore the email? (no)
V	18	1	✓	✓	*In 2009* the heavy storms caused incredible damage and *destroyed* old cottages on the coast.
					Q (task 1): Did the storms cause damage? (yes)
V	19	1	✓	✓	*Yesterday* Rachel and Bryan celebrated outside and *denied* everybody access to their house.
					Q (task 2): Did Bryan and Rachel celebrate in the house? (no)
V	20	1	✓		*The other day* John and his wife noticed the strike message and *worried* at once that their plane might not depart.
V	21	1	✓	✓	*Months ago* Vincent and Frederic cared long for their father and *relied* on their brother to come over as well.
					Q (task 2): Does Vincent have a brother? (yes)
V	22	1	✓	✓	*Back then* Sam and Ron often visited Amy in the office and *argued* with her like crazy.
					Q (task 1): Did Amy visit Sam in the office? (no)
V	23	1	✓		*Two weeks ago* sponsors contributed energy bars and *supplied* athletes with fresh water.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
					Q (task 1): Did sponsors contribute fresh water? (yes)
V	24	1	✓	✓	*The other day* rain and snow prevailed and *followed* us on our hike in the mountains.
					Q (task 2): Did we hike in the mountains? (yes)
V	25	1	✓	✓	*In 2012* the twins travelled Europe and *enjoyed* ancient Roman cities in particular.
					Q (task 2): Did the twins travel to ancient cities? (yes)
V	26	1	✓		*Recently* Cathy's niece and nephew developed a fever and *cried* almost the whole journey back.
V	27	1	✓	✓	*In 2014* Tom and Lara founded a startup and quickly *identified* with their role as their own boss.
					Q (task 1): Did Lara and Tom found a startup? (yes)
V	28	1	✓	✓	*Yesterday* the results appeared online and *satisfied* everyone except for those who had failed.
					Q (task 1): Was everyone happy with the results? (no)
N	1	1	✓	✓	You know, Josh needed *several years* of experience with clients before getting promoted.
					Q (task 1): Did Josh's clients need experience before getting promoted? (no)
N	2	1	✓	✓	As expected, they had *a lot of* urgent *things* on their agenda in the final phase.
					Q (task 1): Were they busy in their final phase? (yes)
N	3	1	✓	✓	In the past, they had *a few problems* with other inhabitants of their building.
					Q (task 2): Do other inhabitants live in their building? (yes)
N	4	1	✓		For the final exam, the professor repeated *all the* important *medical terms* and the contents of six chapters.
N	5	1	✓	✓	My surgeon has operated on *many hands* in his truly exceptional career.
					Q (task 2): Is my surgeon exceptional? (yes)
N	6	1	✓	✓	Regrettably, Tina reached *both members* of the society only after the meeting.
					Q (task 1): Did Tina reach somebody from the society? (yes)
N	7	1	✓		With the loyalty card I got *several points* extra for every item purchased there.
					Q (task 1): Do I have a loyalty card? (yes)
N	8	1	✓	✓	The seminar covered *a lot of areas* of interest for participants from twelve countries.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
					Q (task 2): Was the seminar of interest for participants from different countries? (yes)
N	9	1	✓	✓	Seth left for *a few hours* in order to support his dad with the preparations.
					Q (task 2): Was Seth supported with the preparations? (no)
N	10	1	✓		In fact she searched for *various ways* out of her ever worsening misery.
N	11	1	✓	✓	For their research proposal they investigated *various eyes* with regard to the effect of smoking on sight.
					Q (task 1): Did they investigate the effect of smoking on sight? (yes)
N	12	1	✓	✓	Unexpectedly, Spencer shared *several ideas* about the concept with his colleagues.
					Q (task 1): Did Spencer have a concept in mind? (yes)
N	13	1	✓		Lucas asked the company *a few questions* about the advertisement on the company's website.
					Q (task 1): Was Lucas aware of the advertisement on the company's website? (yes)
N	14	1	✓	✓	Bilateral hearing loss can affect *both ears* at an alarming rate.
					Q (task 2): Can bilateral hearing loss occur at alarming rates? (yes)
N	15	1	✓	✓	Bill told us *many details* about the terrible accident he had been involved in.
					Q (task 2): Had Bill been involved in an accident? (yes)
N	16	1	✓	✓	I just saw *all the* nice *shoes* in the fashion magazine that Beth gave me.
					Q (task 1): Did Beth give me a fashion magazine? (yes)
N	17	1	✓		As you can imagine, I had *many reasons* why I wasn't at the party that Tuesday.
					Q (task 1): Was I at the party? (no)
N	18	1	✓	✓	Ann had to spend *both days* ill in bed because of a severe headache.
					Q (task 1): Did Ann have a headache? (yes)
N	19	1	✓	✓	The campaign addresses *all girls* irrespective of age or nationality.
					Q (task 2): Is the campaign limited to certain age groups? (no)
N	20	1	✓		Francis was persuaded by *a few friends* at university to go to the opera on Friday.
N	21	1	✓	✓	They are hiring *many teachers* in addition to the ones they have.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
					Q (task 2): Do they already have teachers? (yes)
N	22	1	✓	✓	They collaborated with *various schools* in the region to try out different methods.
					Q (task 1): Did they try out different methods? (yes)
N	23	1	✓		They managed to select *various students* on campus for an outstanding team.
					Q (task 1): Did they select someone for the team? (yes)
N	24	1	✓	✓	They engaged *a lot of boys* who helped out in the breaks between the matches.
					Q (task 2): Were there breaks between the matches? (yes)
N	25	1	✓	✓	Jessica and Tony bring along *several games* whenever they are invited over.
					Q (task 2): Are Jessica and Tony invited over sometimes? (yes)
N	26	1	✓		During the exhibition, they reminded *a lot of parents* to let the children sit in the front.
N	27	1	✓	✓	A wrong walking style can harm *both little toes* in particular, as my mother always jokes.
					Q (task 1): Does my mother joke about the harms of a wrong walking style? (yes)
N	28	1	✓	✓	They used special makeup for *all the fingers* and faces to create a spooky Halloween outfit.
					Q (task 1): Did they prepare for Halloween? (yes)
D	1	5	✓		Finn demonstrated his surfing skills and encouraged me to learn *[ ]* as well, but I was too anxious.
					Q (task 1): Did I learn surfing from Finn? (no)
D	2	5	✓		The two will only buy *[ ]* later this week, so they cannot give feedback yet.
D	3	5	✓		Ella got *[ ]* from the butcher at the corner and drops by there frequently.
D	4	6	✓		Emily has been to that stadium on and off, and she likes *[ ]* very much in fact.
					Q (task 1): Has Emily been to that stadium? (yes)
D	5	6	✓		Alex didn't dare to have *[ ]*, so we had to make him rent the suit.
D	6	6	✓		In our baking tutorial she insisted that you must bake *[ ]* because the dough cannot be eaten raw.
					Q (task 1): Did we take a baking tutorial? (yes)
D	7	8	✓		*[ ]* Looked exciting, so Henry will get in touch with them on Thursday morning.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
D	8	7	✓		*[ ]* Didn't shout, but I was aware that my grandma hadn't crossed the street.
					Q (task 1): Had my grandma crossed the street? (no)
D	9	8	✓		They went to the get-together, but then *[ ]* were not <u>self-confident</u> enough to chat there.
					Q (task 1): Did they chat at the get-together? (no)
D	10	7	✓		*[ ]* Have no clue which gym to pick, but I have to do more for my fitness.
					Q (task 1): Have I picked a gym? (no)
D	11	8	✓		Ina arrived early, so *[ ]* <u>waited</u> because she didn't know what else to do.
D	12	7	✓		*[ ]* Shouldn't inform him, but I'm somehow afraid that he could know from elsewhere.
D	13	10	✓		*Although* Mary is not an expert, *but* she's <u>competent</u> and can explain the theory.
					Q (task 1): Is Mary an expert? (no)
D	14	10	✓		*Although* I was tired in the last round, *but* I <u>scored three goals</u>.
D	15	9	✓		*Although* I've communicated in Spanish extensively, *but* I wouldn't consider myself a fluent speaker.
					Q (task 1): Do I consider myself fluent in Spanish? (no)
D	16	10	✓		*Although* Sophie had told me she would be gone by January, *but* I <u>met her in March</u>.
D	17	9	✓		*Although* the tutor began with examples, *but* we didn't understand them.
					Q (task 1): Did the tutor mention examples? (yes)
D	18	9	✓		*Although* the text is not about Ken's core topic, *but* he can refer to it.
D	19	11	✓		Mark claimed he would have an advantage there, but this wasn't the case.
					Q (task 1): Did Mark have an advantage? (no)
D	20	12	✓		Matthew lived in a college hall during his first months of <u>retirement</u>.
D	21	12	✓		Selma didn't approve of the course, because it was really <u>strange</u> compared to the previous one.
					Q (task 1): Did Selma approve of the course? (no)
D	22	11	✓		Kim had planned to accompany her sister, but she had no money to do so.
D	23	12	✓		Erna never got a glimpse of the article, but she made <u>critical comments</u> on the attached sheet of paper.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
					Q (task 1): Did Erna get a glimpse of the article? (no)
D	24	11	✓		The waiters didn't serve the dessert because there was no space on the table.
D	25	12	✓		The deadline had nearly passed, so they hurried to get started.
D	26	11	✓		I have the impression that Bob has great potential to excel in his business.
D	27	12	✓		Those refresher courses are very frustrating because they never comprise outdated stuff.
					Q (task 1): Are the refresher courses frustrating? (yes)
D	28	11	✓		Hannah knew which subjects she wanted to choose, but she had to check whether they clashed.
					Q (task 1): Did Hannah know whether her subjects clashed? (no)
D	29	5		✓	Peter didn't believe that he could do *[ ]*, but Alice reassured him.
D	30	6		✓	Susan was provided a promotion, and she took *[ ]* without hesitating for a second.
D	31	6		✓	Jennifer had the possibility to accept the appointment, so she had to take *[ ]* before another person would.
					Q (task 2): Did Jennifer decline the appointment request? (no)
D	32	5		✓	Tim wondered but he didn't say, and so nobody was irritated.
					Q (task 2): Did Tim wonder? (yes)
D	33	7		✓	*[ ]* Will go to my aunt's birthday on Saturday to have dinner with her.
					Q (task 2): Is my aunt's birthday on Sunday? (no)
D	34	7		✓	*[ ]* Not sure whether my summary had four hundred words, but I did my best.
D	35	8		✓	*[ ]* Should fry it according to the recipe and then put it in the fridge.
D	36	8		✓	*[ ]* Don't recall the names of the guests, but it was a wonderful reunion.
					Q (task 2): Was I at a reunion? (yes)
D	37	10		✓	*Though* Eric hasn't read the abstract, *but* he will be there.
					Q (task 2): Does Eric know the abstract? (no)
D	38	10		✓	*Although* they were quite a big crowd, *but* they didn't push through their suggestion.
D	39	9		✓	*Although* Catherine had a tight schedule, *but* she took time off.

(Continued)

set type	set	condition	task 1	task 2	text (target forms, time adverbials, quantifiers, and critical parts in set type D are in italics; semantically awkward material is underlined)
					Q (task 2): Did Catherine take time off? (yes)
D	40	9		✓	*Although* I'm not that creative, *but* I really benefited from her tips.
D	41	12		✓	Anne was busy, but she had enough resources to write her poem <u>in the attic</u>.
D	42	12		✓	On Sunday evening Joyce and Sally went to the zoo, and they had <u>brunch</u> afterwards.
					Q (task 2): Did Joyce and Sally have brunch before going to the zoo? (no)
D	43	11		✓	Kevin extended his trip to destinations which hadn't been on his initial itinerary.
					Q (task 2): Did Kevin stick to his initial itinerary? (no)
D	44	11		✓	It's fascinating how Stella can fulfil her dream of becoming an actress.
D	45	12		✓	Kathy drove to the mall to get a <u>vehicle</u> for her grandpa's anniversary in June.
					Q (task 2): Is the anniversary of Kathy's grandpa in July? (no)
D	46	12		✓	Grace didn't run the marathon, because her son's babysitter <u>didn't take part either</u>.
D	47	11		✓	Sarah revised her paper during a delayed train ride to her uncle Harry's.
D	48	11		✓	For her weekend Trish wishes she could relax and not be stressed out.
					Q (task 2): Does Trish wish for a relaxing weekend? (yes)
P	1		✓		They contacted Tracy about her subscription, but she didn't answer.
					Q (task 1): Was Tracy contacted about her subscription? (yes)
P	2		✓		They joked that when you go there you crave the whole store.
P	3		✓		Dan was done with his draft, so he corrected it then and there.
					Q (task 1): Did Dan correct his draft? (yes)
P	4			✓	Sandra thought long about the perfect welcome gift for her family.
					Q (task 2): Did Sandra think long about a welcome gift? (yes)
P	5			✓	I'm curious what the bungalow will turn into when Julia is not there any more.
P	6			✓	It doesn't make sense to go there at lunchtime, because it's always full.
					Q (task 2): Does it make sense to go there at lunchtime? (no)

E.4 Results

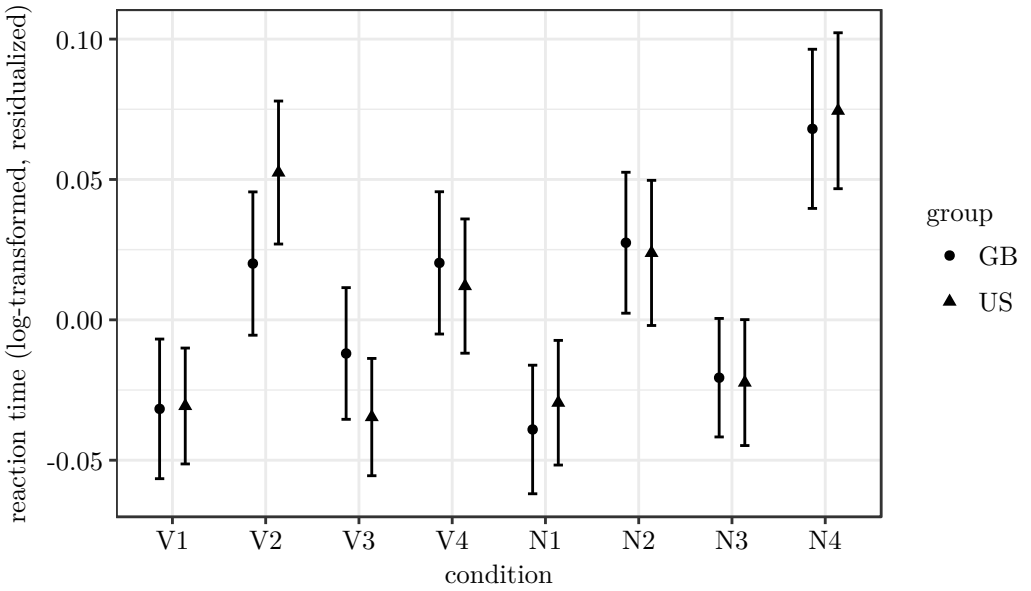

Figure E.4: 95 percent confidence intervals for reaction times by condition and group

Table E.3: Model output (model.spr.pre)

	Estimate	Std. Error	t-value	p-value	
Intercept	−0.0048	0.0199	−0.2420	0.8100	
Condition					
Condition1[V4]	0.0108	0.0129	0.8384	0.4018	
Condition2[N2]	0.0208	0.0128	1.6233	0.1045	
Condition3[V1]	−0.0359	0.0130	−2.7674	0.0057	**
Condition4[N4]	0.0687	0.0129	5.3080	0.0000	***
Condition5[N3]	−0.0230	0.0129	−1.7809	0.0749	
Condition6[V2]	0.0293	0.0130	2.2505	0.0244	*
Condition7[V3]	−0.0306	0.0129	−2.3716	0.0177	*
Group					
GroupHK	−0.0048	0.0099	−0.4856	0.6272	
GroupSG	0.0092	0.0077	1.1857	0.2358	

(Continued)

	Estimate	Std. Error	t-value	p-value	
GroupIN	0.0052	0.0178	0.2921	0.7702	
Stimulus specifics					
Frequency	−0.0003	0.0008	−0.3876	0.6983	
SpellingChangeNoChange	0.0058	0.0110	0.5250	0.5996	
EquivalentNoEquivalent	0.0073	0.0092	0.7957	0.4262	
Speaker background variables					
Age	0.0003	0.0003	1.1270	0.2597	
FirstLanguageChinese	0.0014	0.0092	0.1498	0.8809	
FirstLanguageOther	−0.0170	0.0173	−0.9807	0.3267	
FirstLanguageEnglishChinese	0.0043	0.0144	0.3013	0.7632	
FirstLanguageEnglishOther	0.0314	0.0238	1.3178	0.1876	
FirstLanguageNotAnswered	−0.0520	0.0306	−1.6986	0.0894	
SexMale	0.0045	0.0059	0.7556	0.4499	
EducationBachelor'sDegree	−0.0067	0.0064	−1.0436	0.2967	
EducationNoBa.'sDegree(Yet)	−0.0094	0.0072	−1.3044	0.1921	
EducationNotAnswered	0.0188	0.0308	0.6119	0.5406	
LinguisticsClassesNo	0.0051	0.0066	0.7645	0.4446	
DeviceSmartphone	0.0056	0.0055	1.0180	0.3087	
HandednessLeft	−0.0092	0.0094	−0.9860	0.3241	
Condition:Group					
Condition1[V4]:GroupHK	−0.0261	0.0125	−2.0827	0.0373	*
Condition2[N2]:GroupHK	−0.0039	0.0125	−0.3166	0.7515	
Condition3[V1]:GroupHK	0.0219	0.0125	1.7541	0.0794	
Condition4[N4]:GroupHK	−0.0099	0.0124	−0.7921	0.4283	
Condition5[N3]:GroupHK	0.0332	0.0124	2.6829	0.0073	**
Condition6[V2]:GroupHK	−0.0462	0.0125	−3.7051	0.0002	***
Condition7[V3]:GroupHK	0.0190	0.0125	1.5261	0.1270	
Condition1[V4]:GroupSG	−0.0208	0.0109	−1.9110	0.0560	.
Condition2[N2]:GroupSG	0.0002	0.0108	0.0206	0.9835	
Condition3[V1]:GroupSG	0.0052	0.0109	0.4754	0.6345	
Condition4[N4]:GroupSG	−0.0091	0.0109	−0.8367	0.4027	
Condition5[N3]:GroupSG	0.0270	0.0109	2.4916	0.0127	*

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition6[V2]:GroupSG	−0.0138	0.0109	−1.2605	0.2075	
Condition7[V3]:GroupSG	0.0017	0.0109	0.1595	0.8733	
Condition1[V4]:GroupIN	0.0000	0.0115	0.0026	0.9979	
Condition2[N2]:GroupIN	−0.0126	0.0114	−1.1030	0.2700	
Condition3[V1]:GroupIN	0.0265	0.0115	2.3110	0.0208	*
Condition4[N4]:GroupIN	−0.0326	0.0115	−2.8343	0.0046	**
Condition5[N3]:GroupIN	0.0206	0.0114	1.8010	0.0717	
Condition6[V2]:GroupIN	−0.0272	0.0115	−2.3597	0.0183	*
Condition7[V3]:GroupIN	0.0230	0.0115	1.9997	0.0455	*

Note: *p<0.05; **p<0.01; ***p<0.001

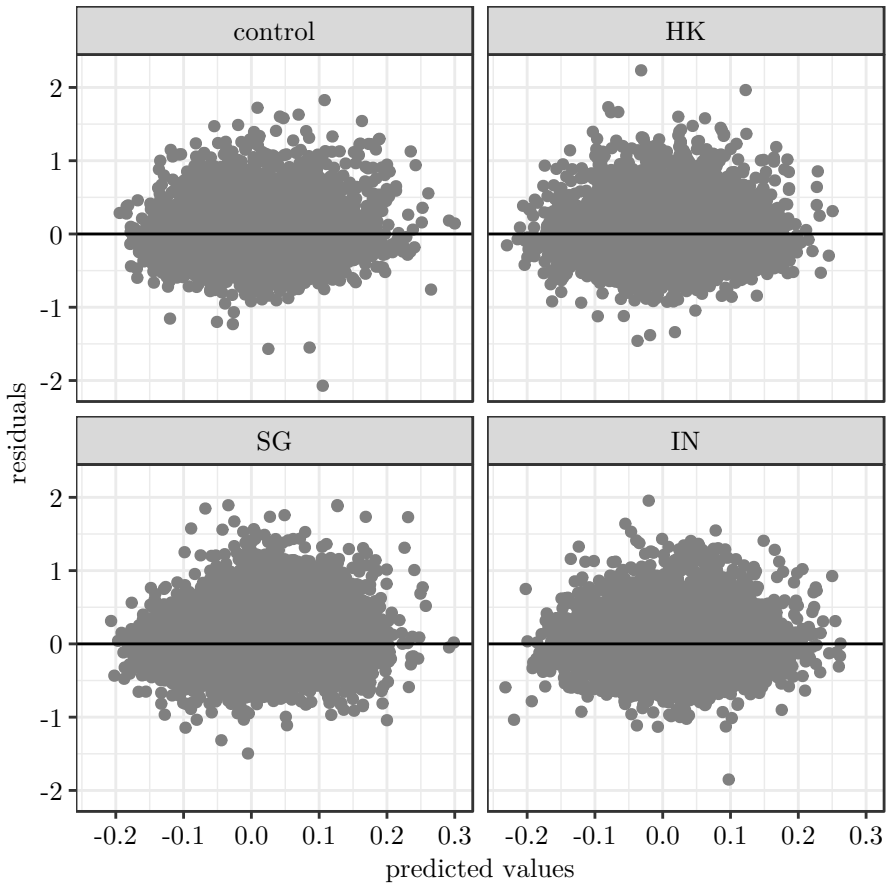

Figure E.5: Residuals by predicted values, by group (model.spr)

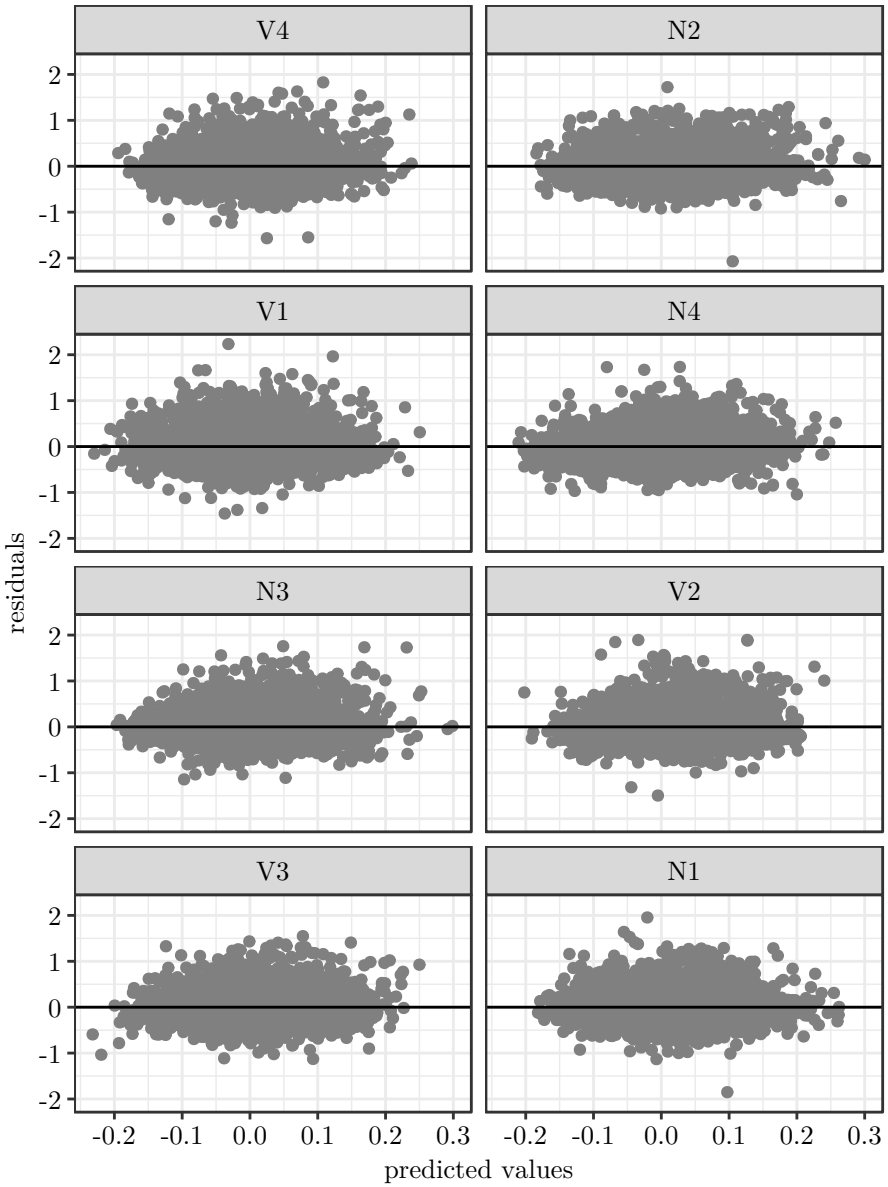

Figure E.6: Residuals by predicted values, by condition (model.spr)

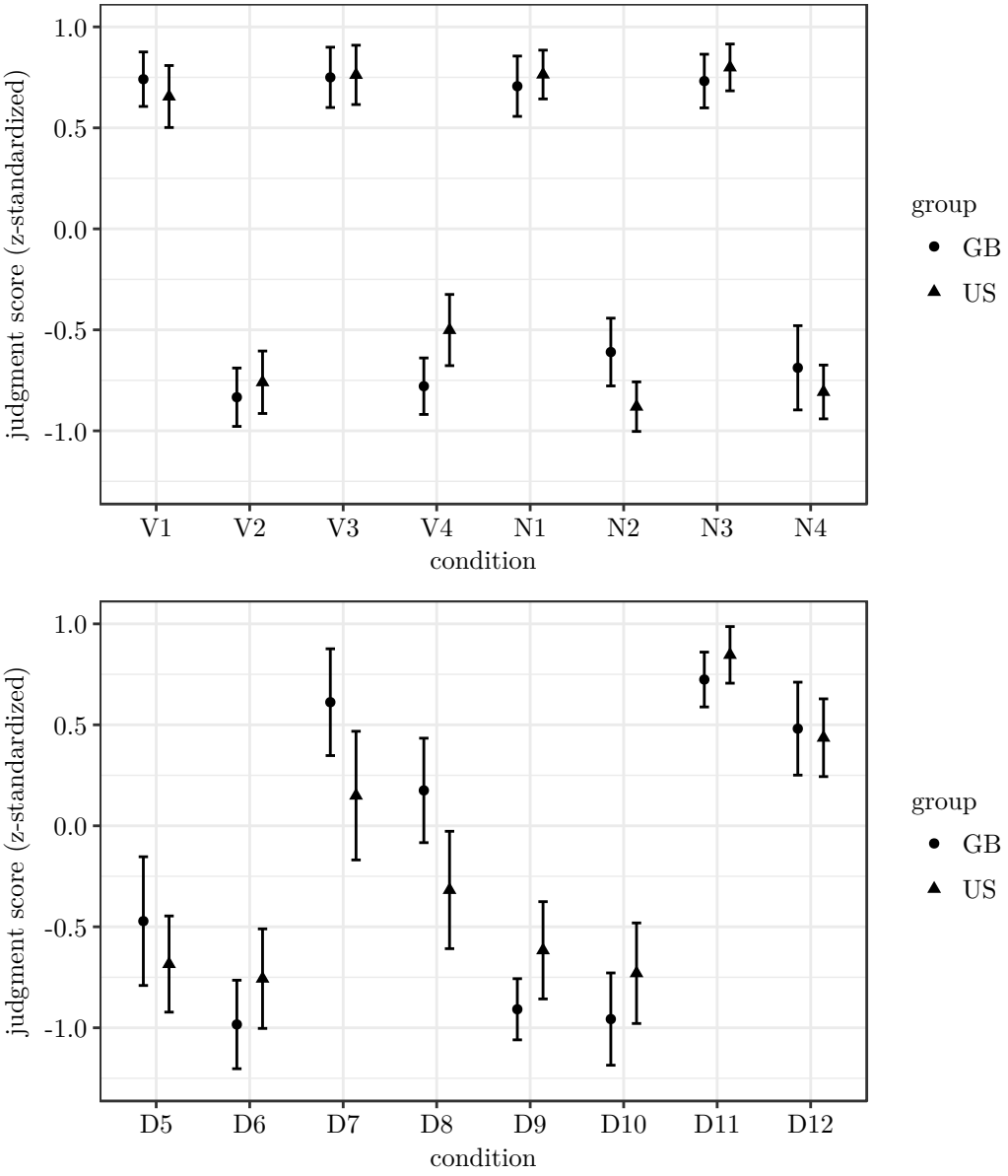

Figure E.7: 95 percent confidence intervals for judgment scores by condition and group

Table E.4: Model output (model.ajt.pre)

	Estimate	Std. Error	t-value	p-value	
Intercept	-0.4597	0.1566	-2.9351	0.0033	**
Condition					
Condition1[V4]	−0.5011	0.0829	−6.0410	0.0000	***
Condition2[D7]	0.2768	0.1879	1.4728	0.1408	
Condition3[D8]	0.1592	0.1846	0.8623	0.3885	
Condition4[D5]	−0.5143	0.1780	−2.8890	0.0039	**
Condition5[D12]	0.6432	0.1323	4.8615	0.0000	***
Condition6[N1]	0.8396	0.0810	10.3645	0.0000	***
Condition7[D9]	−0.7711	0.1820	−4.2375	0.0000	***
Condition8[D6]	−0.8699	0.1821	−4.7777	0.0000	***
Condition9[V3]	0.8916	0.0821	10.8616	0.0000	***
Condition10[V2]	−0.6642	0.0839	−7.9148	0.0000	***
Condition11[N3]	0.8770	0.0810	10.8251	0.0000	***
Condition12[D10]	−0.8561	0.1856	−4.6136	0.0000	***
Condition13[N4]	−0.6477	0.0815	−7.9461	0.0000	***
Condition14[V1]	0.8256	0.0824	10.0193	0.0000	***
Condition15[N2]	−0.6531	0.0809	−8.0754	0.0000	***
Group					
GroupHK	0.0037	0.0362	0.1008	0.9197	
GroupSG	0.0319	0.0286	1.1151	0.2648	
GroupIN	0.0994	0.0593	1.6766	0.0936	
Stimulus specifics					
Frequency	0.0379	0.0138	2.7447	0.0061	**
Speaker background variables					
Age	−0.0003	0.0009	−0.2924	0.7700	
FirstLanguageChinese	0.0012	0.0305	0.0392	0.9687	
FirstLanguageOther	−0.0247	0.0554	−0.4458	0.6557	
FirstLanguageEnglishChinese	0.0038	0.0441	0.0859	0.9315	

(Continued)

	Estimate	Std. Error	t-value	p-value	
FirstLanguageEnglishOther	−0.0101	0.0858	−0.1177	0.9063	
FirstLanguageNotAnswered	−0.0677	0.1003	−0.6751	0.4996	
SexMale	0.0017	0.0199	0.0859	0.9316	
EducationBachelor'sDegree	−0.0008	0.0218	−0.0389	0.9690	
EducationNoBa.'sDegree(Yet)	0.0028	0.0254	0.1119	0.9109	
EducationNotAnswered	0.0441	0.0947	0.4654	0.6416	
LinguisticsClassesNo	0.0040	0.0226	0.1777	0.8589	
Condition:Group					
Condition1[V4]:GroupHK	0.3303	0.0959	3.4450	0.0006	***
Condition2[D7]:GroupHK	0.1597	0.1658	0.9632	0.3355	
Condition3[D8]:GroupHK	−0.1074	0.1466	−0.7325	0.4639	
Condition4[D5]:GroupHK	0.3047	0.1437	2.1205	0.0340	*
Condition5[D12]:GroupHK	−0.4553	0.1062	−4.2889	0.0000	***
Condition6[N1]:GroupHK	−0.1912	0.0940	−2.0339	0.0420	*
Condition7[D9]:GroupHK	−0.2767	0.1434	−1.9294	0.0537	.
Condition8[D6]:GroupHK	0.1079	0.1486	0.7258	0.4679	
Condition9[V3]:GroupHK	−0.1221	0.0950	−1.2855	0.1986	
Condition10[V2]:GroupHK	0.3004	0.0962	3.1229	0.0018	**
Condition11[N3]:GroupHK	−0.2985	0.0946	−3.1569	0.0016	**
Condition12[D10]:GroupHK	−0.0419	0.1517	−0.2759	0.7826	
Condition13[N4]:GroupHK	0.3245	0.0950	3.4155	0.0006	***
Condition14[V1]:GroupHK	−0.2288	0.0956	−2.3928	0.0167	*
Condition15[N2]:GroupHK	0.3624	0.0946	3.8321	0.0001	***
Condition1[V4]:GroupSG	−0.0159	0.0794	−0.2003	0.8412	
Condition2[D7]:GroupSG	−0.0543	0.1394	−0.3896	0.6968	
Condition3[D8]:GroupSG	0.0493	0.1202	0.4102	0.6817	
Condition4[D5]:GroupSG	0.0448	0.1197	0.3743	0.7082	
Condition5[D12]:GroupSG	−0.1448	0.0895	−1.6180	0.1057	
Condition6[N1]:GroupSG	−0.1404	0.0783	−1.7938	0.0728	
Condition7[D9]:GroupSG	0.4497	0.1175	3.8278	0.0001	***
Condition8[D6]:GroupSG	0.1151	0.1210	0.9507	0.3418	
Condition9[V3]:GroupSG	−0.0720	0.0787	−0.9154	0.3600	

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition10[V2]:GroupSG	0.2405	0.0807	2.9793	0.0029	**
Condition11[N3]:GroupSG	−0.2350	0.0783	−2.9996	0.0027	**
Condition12[D10]:GroupSG	0.0682	0.1265	0.5390	0.5899	
Condition13[N4]:GroupSG	0.1021	0.0790	1.2931	0.1960	
Condition14[V1]:GroupSG	−0.3302	0.0794	−4.1601	0.0000	***
Condition15[N2]:GroupSG	0.0661	0.0782	0.8451	0.3981	
Condition1[V4]:GroupIN	−0.1523	0.0870	−1.7498	0.0802	
Condition2[D7]:GroupIN	−0.0089	0.1432	−0.0622	0.9504	
Condition3[D8]:GroupIN	0.2133	0.1282	1.6633	0.0963	
Condition4[D5]:GroupIN	0.4642	0.1266	3.6676	0.0002	***
Condition5[D12]:GroupIN	−0.3545	0.0960	−3.6930	0.0002	***
Condition6[N1]:GroupIN	−0.3964	0.0849	−4.6676	0.0000	***
Condition7[D9]:GroupIN	0.6742	0.1268	5.3155	0.0000	***
Condition8[D6]:GroupIN	0.3638	0.1305	2.7877	0.0053	**
Condition9[V3]:GroupIN	−0.3341	0.0843	−3.9637	0.0001	***
Condition10[V2]:GroupIN	−0.0115	0.0867	−0.1324	0.8947	
Condition11[N3]:GroupIN	−0.3914	0.0839	−4.6665	0.0000	***
Condition12[D10]:GroupIN	0.2328	0.1341	1.7353	0.0827	
Condition13[N4]:GroupIN	0.3796	0.0860	4.4154	0.0000	***
Condition14[V1]:GroupIN	−0.3875	0.0856	−4.5283	0.0000	***
Condition15[N2]:GroupIN	0.2764	0.0850	3.2523	0.0011	**

Note: *p<0.05; **p<0.01; ***p<0.001

Table E.5: Model output (model.ajt.vn)

	Estimate	Std. Error	t-value	p-value	
Intercept	−0.5008	0.1632	−3.0686	0.0022	**
Condition					
Condition1[V4]	−0.6161	0.0736	−8.3665	0.0000	**
Condition2[N2]	−0.7796	0.0725	−10.7499	0.0000	**
Condition3[V1]	0.7110	0.0732	9.7157	0.0000	**

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition4[N4]	−0.7729	0.0731	−10.5716	0.0000	**
Condition5[N3]	0.7492	0.0727	10.3032	0.0000	**
Condition6[V2]	−0.7791	0.0745	−10.4540	0.0000	**
Condition7[V3]	0.7753	0.0729	10.6364	0.0000	**
Group					
GroupHK	0.0627	0.0330	1.8990	0.0576	.
GroupSG	−0.0141	0.0273	−0.5172	0.6050	
GroupIN	−0.0455	0.0295	−1.5418	0.1231	
Stimulus specifics					
LogLemmaFreq	0.0529	0.0169	3.1205	0.0018	**
Condition:Group					
Condition1[V4]:GroupHK	0.2694	0.0880	3.0598	0.0022	**
Condition2[N2]:GroupHK	0.3045	0.0869	3.5053	0.0005	***
Condition3[V1]:GroupHK	−0.2901	0.0878	−3.3039	0.0010	***
Condition4[N4]:GroupHK	0.2664	0.0872	3.0536	0.0023	**
Condition5[N3]:GroupHK	−0.3586	0.0869	−4.1279	0.0000	***
Condition6[V2]:GroupHK	0.2411	0.0883	2.7299	0.0063	**
Condition7[V3]:GroupHK	−0.1823	0.0873	−2.0895	0.0367	*
Condition1[V4]:GroupSG	0.0478	0.0726	0.6578	0.5107	
Condition2[N2]:GroupSG	0.1244	0.0716	1.7369	0.0824	
Condition3[V1]:GroupSG	−0.2800	0.0726	−3.8557	0.0001	***
Condition4[N4]:GroupSG	0.1551	0.0723	2.1458	0.0319	*
Condition5[N3]:GroupSG	−0.1922	0.0718	−2.6746	0.0075	**
Condition6[V2]:GroupSG	0.2799	0.0739	3.7864	0.0002	***
Condition7[V3]:GroupSG	−0.0376	0.0721	−0.5219	0.6017	
Condition1[V4]:GroupIN	−0.0296	0.0795	−0.3715	0.7103	
Condition2[N2]:GroupIN	0.3991	0.0778	5.1326	0.0000	***
Condition3[V1]:GroupIN	−0.2499	0.0783	−3.1927	0.0014	**
Condition4[N4]:GroupIN	0.5299	0.0785	6.7467	0.0000	***
Condition5[N3]:GroupIN	−0.2811	0.0767	−3.6635	0.0002	***

(Continued)

	Estimate	Std. Error	t-value	p-value	
Condition6[V2]:GroupIN	0.1186	0.0793	1.4952	0.1349	
Condition7[V3]:GroupIN	−0.2141	0.0771	−2.7771	0.0155	*

Note: *p<0.05; **p<0.01; ***p<0.001

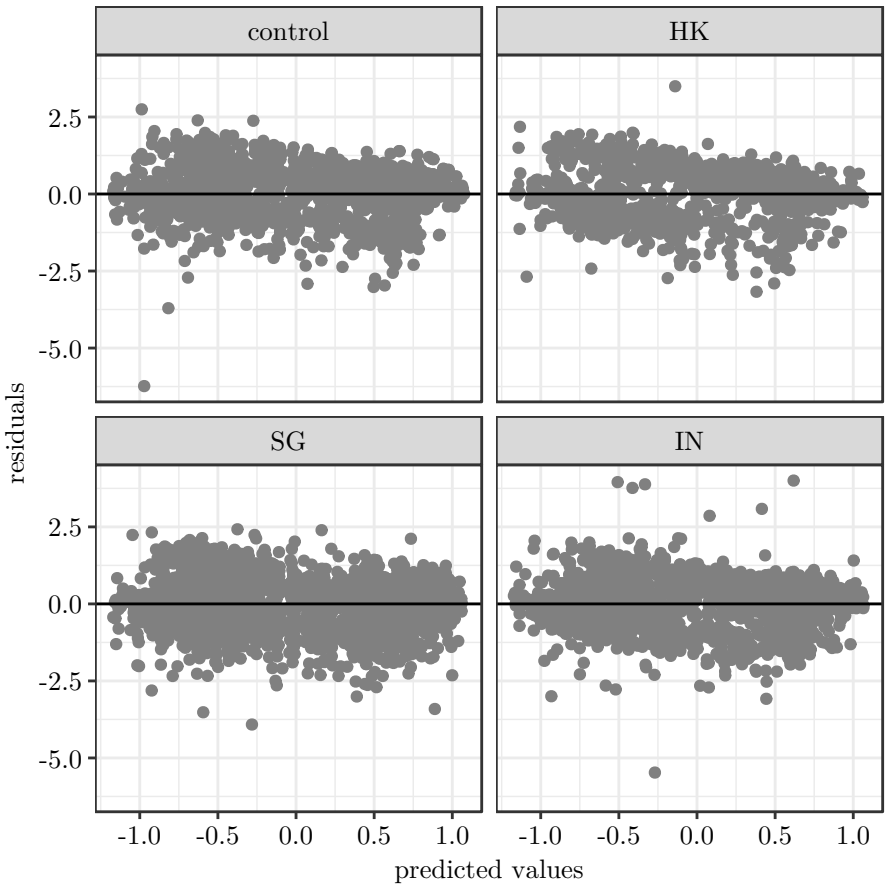

Figure E.8: Residuals by predicted values, by group (model.ajt)

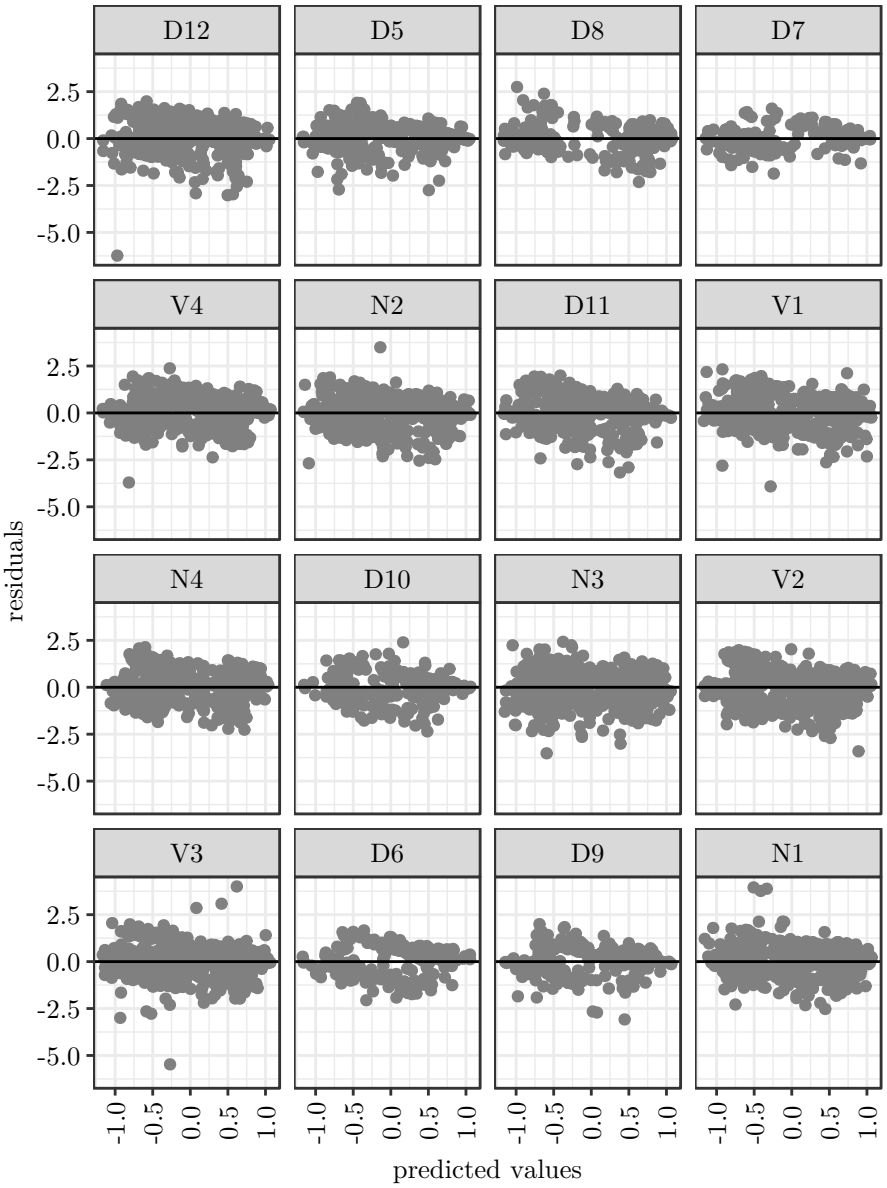

Figure E.9: Residuals by predicted values, by condition (model.ajt)

List of references

Alsagoff, Lubna. 2001. Tense and aspect in Singapore English. In Vincent B. Y. Ooi (ed.), *Evolving identities: The English language in Singapore and Malaysia*, 79–88. Singapore: Times Academic Press.

Alsagoff, Lubna & Chee Lick Ho. 1998. The grammar of Singapore English. In Joseph A. Foley, Thiru Kandiah, Zhiming Bao, Anthea Fraser Gupta, Lubna Alsagoff, Chee Lick Ho, Lionel Wee, Ismail S. Talib & Wendy Bokhorst-Heng (eds.), *English in new cultural contexts: Reflections from Singapore*, 127–151. Singapore: Singapore Institute of Management.

Ansaldo, Umberto. 2004. The evolution of Singapore English: Finding the matrix. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 127–149. Amsterdam & Philadelphia: John Benjamins.

Anthony, Laurence. 2014. *Laurence Anthony's Website: AntConc Homepage*. Version 3.4.3m. http://www.laurenceanthony.net/. Tokyo: Waseda University.

ARCHER-3.2 (Lancaster). 2013. *A Representative Corpus of Historical English Registers*. Version 3.2. 1990-2013.

Baayen, R. Harald. 2012. *Analyzing linguistic data: A practical introduction to statistics using R*. 6th print. Cambridge: Cambridge University Press.

Bailey, Richard W. & Manfred Görlach. 1982. *English as a world language*. 1st print. Ann Arbor, MI: University of Michigan Press.

Baker, Sidney J. 1945. *The Australian language*. Sydney: Angus & Robertson Ltd. 1978. 3rd edn. Milson's Point, NSW: Currawong Press.

Bamgbose, Ayo. 1998. Torn between the norms: Innovations in world Englishes. *World Englishes* 17(1). 1–14.

Bao, Zhiming. 1995. *Already* in Singapore English. *World Englishes* 14(2). 181–188.

Bao, Zhiming. 1998. The sounds of Singapore English. In Joseph A. Foley, Thiru Kandiah, Zhiming Bao, Anthea Fraser Gupta, Lubna Alsagoff, Chee Lick Ho, Lionel Wee, Ismail S. Talib & Wendy Bokhorst-Heng (eds.), *English in new cultural contexts: Reflections from Singapore*, 152–174. Singapore: Singapore Institute of Management.

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. 2010. A usage-based approach to substratum transfer: The case of four unproductive features in Singapore English. *Language: Journal of the Linguistic Society of America* 86(4). 792–820.

Bao, Zhiming. 2015. *The making of vernacular Singapore English: System, transfer, and filter*. Cambridge: Cambridge University Press.

Bao, Zhiming. 2017. Transfer is transfer; grammaticalization is grammaticalization. In Markku Filppula, Juhani Klemola, Anna Mauranen & Svetlana Vetchinnikova (eds.), *Changing English: Global and local perspectives*, 311–329. Berlin & Boston: De Gruyter Mouton.

Barlow, Michael & Suzanne Kemmer (eds.). 2000. *Usage-based models of language*. Stanford, CA: Center for the Study of Language & Information.

Baroni, Marco & Silvia Bernardini (eds.). 2006. *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Bates, Douglas, Martin Maechler, Ben Bolker & Steven Walker. 2017. *Package 'lme4'*. https://cran.r-project.org/web/packages/lme4/lme4.pdf.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. 1st publ. Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora: Corpus-Based Language Learning, Language Processing and Linguistics* 10(1). 11–45.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. 1st publ. Harlow: Longman.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 2007. *Longman grammar of spoken and written English*. 7th impr. Harlow: Longman.

Biewer, Carolin. 2011. Modal auxiliaries in second language varieties of English: A learner's perspective. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 7–34. Amsterdam: John Benjamins.

Biewer, Carolin. 2015. *South Pacific Englishes: A sociolinguistic and morphosyntactic profile of Fiji English, Samoan English and Cook Islands English*. Amsterdam & Philadelphia: John Benjamins.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge: MIT Press.

Bolton, Kingsley (ed.). 2000. *Hong Kong English: Autonomy and creativity*. Special Issue of World Englishes 19(3). Published as a book 2002. Hong Kong: Hong Kong University Press.

Bolton, Kingsley (ed.). 2002a. *Hong Kong English: Autonomy and creativity*. Hong Kong: Hong Kong University Press.

Bolton, Kingsley. 2002b. The sociolinguistics of Hong Kong and the space for Hong Kong English. In Kingsley Bolton (ed.), *Hong Kong English: Autonomy and creativity*, 29–55. Hong Kong: Hong Kong University Press.

Bolton, Kingsley. 2003. *Chinese Englishes: A sociolinguistic history*. Cambridge: Cambridge University Press.

Bolton, Kingsley. 2006. World Englishes today. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The handbook of world Englishes*, 240–270. Malden, MA: Wiley-Blackwell.

Bolton, Kingsley & Helen Kwok. 1990. The dynamics of the Hong Kong accent: Social identity and sociolinguistic description. *Journal of Asian Pacific Communication* 1(1). 147–172.

Bolton, Kingsley & Shirley G.-l. Lim. 2002. Futures for Hong Kong English. In Kingsley Bolton (ed.), *Hong Kong English: Autonomy and creativity*, 295–313. Hong Kong: Hong Kong University Press.

Bongartz, Christiane M. & Sarah Buschfeld. 2011. English in Cyprus: Second language variety or learner English? In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 35–54. Amsterdam: John Benjamins.

Bruthiaux, Paul. 2003. Squaring the circles: Issues in modeling English worldwide. *International Journal of Applied Linguistics* 13(2). 159–178.

Bryant, David & Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255–265.

Budge, Carol. 1989. Plural marking in Hong Kong English. *Hongkong Papers in Linguistics and Language Teaching* 12. 39–48.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 65–85. Stanford: Center for the Study of Language & Information.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan L. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3). 261–290.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language: Journal of the Linguistic Society of America* 82(4). 711–733.

Bybee, Joan L. (ed.). 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, Joan L. & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language: Journal of the Linguistic Society of America* 82(2). 323–355.

Bybee, Joan L. & Paul J. Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, Joan L. & Dan I. Slobin. 1982. Rules and schemes in the development and use of the English past tense. *Language: Journal of the Linguistic Society of America* 58(2). 265–289.

Bybee, Joan L. & Sandra Thompson. 2007. Three frequency effects in syntax. In Joan L. Bybee (ed.), *Frequency of use and the organization of language*, 269–278. Oxford: Oxford University Press.

Carls, Uwe. 1999. Compounding in Indian English. In Uwe Carls & Peter Lucko (eds.), *Form, function, and variation in English: Studies in honour of Klaus Hansen*, 141–153. Frankfurt: Peter Lang.

Census and Statistics Department Hong Kong. 2016. *Thematic Household Survey Report No. 59*. https://www.statistics.gov.hk/pub/B11302592016XXXXB0100.pdf (8 February, 2017).

Census and Statistics Department Hong Kong. 2017. *2016 Hong Kong Population By-census*. https://www.bycensus2016.gov.hk/data/16bc-summary-results.pdf (16 September, 2018).

Chambers, Jack K. 2001. Vernacular universals. In Josep M. Fontana, Louise McNally, M. Teresa Turell & Enric Vallduví (eds.), *ICLaVE 1: Proceedings of the First International Conference on Language Variation in Europe*, 52–60. Barcelona: Universitat Pompeu Fabra.

Chambers, Jack K. 2004. Dynamic typology and vernacular universals. In Bernd Kortmann (ed.), *Dialectology meets typology: Dialect grammar from a cross-linguistic perspective*, 127–145. Berlin: Mouton de Gruyter.

Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Cheng, Winnie, Chris Greaves & Martin Warren. 2008. *A corpus-driven study of discourse intonation: The Hong Kong corpus of spoken English (prosodic)*. Amsterdam & Philadelphia: John Benjamins.

Chng, Huang Hoon. 2003. "You See Me No Up": Is Singlish a problem? *Language Problems and Language Planning* 27(1). 45–62.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press.

Constitution. 1965. *Singapore Statutes Online: Constitution of the Republic of Singapore*. https://sso.agc.gov.sg/Act/CONS1963 (3 October, 2018).

Crewe, William (ed.). 1977. *The English language in Singapore*. Singapore: Eastern University Press.

Croft, William. 2000. *Explaining language change: An evolutionary approach*. Harlow: Longman.

Croft, William & David A. Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.

Crump, Matthew J. C., John V. McDonnell & Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE* 8(3). 1–18.

Dalal, Dev K. & Michael J. Zickar. 2012. Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods* 15(3). 339–362.

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. http://corpus.byu.edu/coca/ (24 March, 2015).

Davies, Mark. 2013. *Corpus of Gobal Web-Based English (GloWbE)*. http://corpus.byu.edu/glowbe/ (29 July, 2014).

Davies, Mark & Robert Fuchs. 2015. 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). 1–47.

Davydova, Julia. 2011. *The present perfect in non-native Englishes, a corpus-based study of variation*. Berlin & Boston: De Gruyter Mouton.

Department of Statistics Singapore. 1980. *Census of Population 1980*.

Department of Statistics Singapore. 1990. *Census of Population 1990*.

Department of Statistics Singapore. 2000. *Singapore Census of Population 2000, Statistical Release 1: Demographic Characteristics*. https://www.singstat.gov.sg/publications/cop2000/census_stat_release1 (3 October, 2018).

Department of Statistics Singapore. 2010. *Singapore Census of Population 2010, Statistical Release 1: Demographic Characteristics, Education, Language and Religion*. https://www.singstat.gov.sg/-/media/files/publications/cop2010/census_2010_release1/cop2010sr1.pdf (3 October, 2018).

Department of Statistics Singapore. 2015. *General Household Survey 2015*. https://www.singstat.gov.sg/-/media/files/publications/ghs/ghs2015/ghs2015.pdf (3 October, 2018).

Deshors, Sandra C. 2018. Modeling World Englishes in the 21st century. In Sandra C. Deshors (ed.), *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, 1–14. Amsterdam: John Benjamins.

Deterding, David. 2007. *Singapore English*. Edinburgh: Edinburgh University Press.

Deterding, David & Ee Ling Low. 2001. The NIE Corpus of Spoken Singapore English (NIECSSE). *SAAL Quarterly* 56. 2–5.

Diessel, Holger. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25(2). 108–127.

Dijkstra, Ton & Walter J. B. Van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition* 5(3). 175–197.

Dress, Andreas W. M. & Daniel H. Huson. 2004. Constructing splits graphs. *IEEE/ ACM Transactions on Computational Biology and Bioinformatics* 1(3). 109–115.

Eckert, Penelope & Sally McConnell-Ginet. 2013. *Language and gender*. 2nd edn. Cambridge: Cambridge University Press.

Edwards, Alison. 2014. The progressive aspect in the Netherlands and the ESL/EFL continuum. *World Englishes* 33(2). 173–194.

Edwards, Alison. 2016. *English in the Netherlands: Functions, forms and attitudes*. Amsterdam & Philadelphia: John Benjamins.

Elliott, Annie B. 1983. *Errors in English*. Singapore University Press, National University of Singapore.

Ellis, Rod. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Enochson, Kelly & Jennifer Culbertson. 2015. Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLoS ONE* 10(3).

Evans, Stephen. 2011. Hong Kong English: The growing pains of a new variety. *Asian Englishes* 14(1). 22–45.

Evans, Stephen. 2014. The evolutionary dynamics of postcolonial Englishes: A Hong Kong case study. *Journal of Sociolinguistics* 18(5). 571–603.

Field, Andy, Jeremy Miles & Zoë Field. 2012. *Discovering statistics using R*. London: SAGE.

Finkbeiner, Matthew, Tamar H. Gollan & Alfonso Caramazza. 2006. Lexical access in bilingual speakers: What's the (hard) problem? *Bilingualism: Language and Cognition* 9(2). 153–166.

Fletcher, William H. 2004. Making the web more useful as a source for linguistic corpora. In Ulla Connor & Thomas A. Upton (eds.), *Corpus Linguistics in North America 2002. Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*. Amsterdam: Rodopi.

Fletcher, William H. 2007. Concordancing the web: Promise and problems, tools and techniques. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 25–46. Amsterdam & New York: Rodopi.

Foley, Joseph A., Thiru Kandiah, Zhiming Bao, Anthea Fraser Gupta, Lubna Alsagoff, Chee Lick Ho, Lionel Wee, Ismail S. Talib & Wendy Bokhorst-Heng (eds.). 1998. *English in new cultural contexts: Reflections from Singapore*. Singapore: Singapore Institute of Management.

Fong, Vivienne. 2004. The verbal cluster. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 75–104. Amsterdam & Philadelphia: John Benjamins.

Francis, W. Nelson & Henry Kučera (eds.). 1964. *A standard corpus of present-day edited American English, for use with digital computers (Brown)*. Providence, Rhode Island: Brown University.

Gargesh, Ravinder. 2004. Indian English: Phonology. In Edgar W. Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English. Volume 1: Phonology*, 992–1002. Berlin & New York: Mouton de Gruyter.

Gatto, Maristella. 2014. *Web as corpus: Theory and practice*. London: Bloomsbury.

Gil, David. 2013. Numeral classifiers. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Gisborne, Nikolas. 2009. Aspects of the morphosyntactic typology of Hong Kong English. *English World-Wide* 30(2). 149–169.

Government of India, Ministry of Human Resource Development. 2016. *Language Education*. http://mhrd.gov.in/language-education (16 September, 2018).

Granena, Gisela, Daniel O. Jackson & Yucel Yilmaz. 2016. *Cognitive individual differences in second language processing and acquisition*. Amsterdam & Philadelphia: John Benjamins.

Greenbaum, Sidney. 1996. *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, Stefan Th. 2013. *Statistics for linguistics with R: A practical introduction*. 2nd rev. edn. Berlin: De Gruyter Mouton.

Gut, Ulrike. 2005. The realisation of final plosives in Singapore English: Phonological rules & ethnic differences. In David Deterding, Adam Brown & Ee Ling Low (eds.), *English in Singapore: Phonetic research on a corpus*, 14–25. Singapore: McGraw-Hill.

Gut, Ulrike. 2009a. *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt: Peter Lang.

Gut, Ulrike. 2009b. Past tense marking in Singapore English verbs. *English World-Wide* 30(3). 262–277.

Haiman, John. 1994. Ritualization and the development of language. In William Pagliuca (ed.), *Perspectives on grammaticalization*, 3–28. Amsterdam & Philadelphia: John Benjamins.

Hansen, Beke. 2018. *Corpus linguistics and sociolinguistics: A study of variation and change in the modal systems of World Englishes*. Leiden & Boston: Brill Rodopi.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hernández, Nuria. 2006. *User's guide to FRED*. FRED: Freiburg English Dialect Corpus. http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/ (20 May, 2017).

Ho, Mian Lian. 1981. *The noun phrase in Singapore English*. Melbourne: Monash University. MA thesis.

Ho, Mian Lian. 2003. Past tense marking in Singapore English. In David Deterding, Adam Brown & Ee Ling Low (eds.), *English in Singapore: Research on grammar*, 39–47. Singapore: McGraw-Hill.

Ho, Mian Lian & John T. Platt. 1993. *Dynamics of a contact continuum: Singaporean English*. Oxford: Clarendon Press.

Hockett, Charles F. 1958. *A course in modern linguistics*. New York: Macmillan.

Hong Kong Polytechnic University. 2018. *Hong Kong Corpus of Spoken English*. http://rcpce.engl.polyu.edu.hk/HKCSE/ (11 September, 2018).

Hooper, Joan L. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William M. Christie (ed.), *Current Progress in Historical Linguistics*, 96–105. Amsterdam: North-Holland.

Hopper, Paul J. 1979. Aspect and foregrounding in discourse. In Talmy Givón (ed.), *Discourse and syntax*, 213–241. New York: Academic Press.

Horch, Stephanie. 2017. *Conversion in Asian Englishes: A usage-based account of the emergence of new local norms*. Freiburg: Albert-Ludwigs-Universität, Universitätsbibliothek.

Hornby, Albert S. 2005. *Oxford Advanced Learner's Dictionary*. 7th edn. Oxford: Oxford University Press.

Hosali, Priya. 2008. Butler English: Morphology and syntax. In Rajend Mesthrie (ed.), *Varieties of English, 4: Africa, South and Southeast Asia*, 563–577. Berlin: Mouton de Gruyter.

Hsu, Yun Chiao. 1950. An analysis of the Chinese population in Malaya. *Journal of the South Seas Society* 6. 64–74.

Hu, Jianhua, Haihua Pan & Liejiong Xu. 2001. Is there a finite vs. nonfinite distinction in Chinese? *Linguistics* 39(6). 1117–1148.

Huang, C.-T. James. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15(4). 531–574.

Huang, C.-T. James. 1987. Remarks on empty categories in Chinese. *Linguistic Inquiry* 18(2). 321–337.

Huang, C.-T. James. 1989. Pro-drop in Chinese: A generalized control theory. In Osvaldo Jaeggli & Kenneth Safir (eds.), *The null subject parameter*, 185–214. Dordrecht: Kluwer.

Huang, C.-T. James. 1998. *Logical relations in Chinese and the theory of grammar*. New York: Garland.

Hundt, Marianne & Joybrato Mukherjee. 2011. Introduction: Bridging a paradigm gap. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 1–6. Amsterdam: John Benjamins.

Hundt, Marianne, Nadja Nesselhauf & Carolin Biewer (eds.). 2007a. *Corpus linguistics and the web*. Amsterdam & New York: Rodopi.

Hundt, Marianne, Nadja Nesselhauf & Carolin Biewer. 2007b. Corpus linguistics and the web. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 1–6. Amsterdam & New York: Rodopi.

Hung, Tony T. N. 2000. Towards a phonology of Hong Kong English. *World Englishes* 19(3). 337–356.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267.

Ipeirotis, Panagiotis G. 2010. *Demographics of Mechanical Turk*. Working Paper.

Jacoby, William G. 2000. Loess: A nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies* 19. 577–613.

Jaeger, Florian. 2008. *Modeling self-paced reading data: Effects of word length, word position, spill-over, etc.* HLP/Jaeger lab blog. https://hlplab.wordpress.com/2008/01/23/modeling-self-paced-reading-data-effects-of-word-length-word-position-spill-over-etc/ (1 December, 2016).

Jiang, Nan. 2012. *Conducting reaction time research in second language studies*. Routledge: New York.

Johnson, Keith. 1996. Speech perception without speech normalization. In Keith Johnson & John W. Mullennix (eds.), *Talker variability in speech processing*, 145–165. San Diego, CA: Academic Press.

Joseph, John E. 2004. *Language and identity: National, ethnic, religious*. 1st publ. Basingstoke, U.K.: Palgrave Macmillan.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: John Benjamins.

Just, Marcel A. & Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87(4). 329–354.

Just, Marcel A., Patricia A. Carpenter & Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* 111(2). 228–238.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the world: Teaching and learning the language and literatures*, 11–30. Cambridge: Cambridge University Press.

Kachru, Braj B. 1986. *The alchemy of English: The spread, functions, and models of non-native Englishes*. Chicago, IL: University of Illinois Press.

Kachru, Braj B. 1988. The sacred cows of English. *English Today* 4(4). 3–8.

Kachru, Braj B. 1992. Introduction: The other side of English and the 1990s. In Braj B. Kachru (ed.), *The other tongue: English across cultures*, 2nd edn., 1–15. Chicago, IL: University of Illinois Press.

Kachru, Braj B. 2005. *Asian Englishes: Beyond the Canon*. Hong Kong: Hong Kong University Press.

Kachru, Yamuna. 2006. *Hindi*. Amsterdam: John Benjamins.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo & Martin Corley. 2009. Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods* 41(1). 1–12.

Khan, Farhat. 1991. Final consonant cluster simplification in a variety of Indian English. In Jenny Cheshire (ed.), *English around the world: Sociolinguistic perspectives*, 1st publ., 288–298. Cambridge: Cambridge University Press.

Kilgarriff, Adam. 2001. Web as corpus. *Proceedings of the Corpus Linguistics Conference (CL 2001)*. University Centre for Computer Research on Language Technical Paper, Vol. 13, Special Issue, Lancaster University. 342–344.

Kirkici, Bilal. 2010. Distinct mechanisms in the processing of English past tense morphology: A view from L2 processing. In Martin Pütz & Laura Sicola (eds.),

Cognitive processing in second language acquisition: Inside the learner's mind, 67–83. Amsterdam & Philadelphia: John Benjamins.

Kortmann, Bernd (ed.). 2012. *The Mouton World Atlas of Variation in English*. Berlin: De Gruyter Mouton.

Kortmann, Bernd & Kerstin Lunkenheimer. 2013. *The Electronic World Atlas of Varieties of English*. http://ewave-atlas.org/ (17 January, 2015).

Kortmann, Bernd & Benedikt Szmrecsanyi. 2009. World Englishes between simplification and complexification. In Thomas Hoffmann, Lucia Siebers & Edgar W. Schneider (eds.), *World Englishes: Problems, properties and prospects*, 265–285. Amsterdam: John Benjamins.

Kortmann, Bernd & Benedikt Szmrecsanyi (eds.). 2012. *Linguistic complexity: Second language acquisition, indigenization, contact*. Berlin & Boston: De Gruyter.

Kortmann, Bernd & Christoph Wolk. 2012. Morphosyntactic variation in the anglophone world: A global perspective. In Bernd Kortmann (ed.), *The Mouton World Atlas of Variation in English*, 906–936. Berlin: De Gruyter Mouton.

Kumpf, Lorraine. 1984. Temporal systems and universality in interlanguage: A case study. In Fred R. Eckman, Lawrence H. Bell & Diane Nelson (eds.), *Universals of second language acquisition*. Rowley, MA: Newbury House.

Labov, William. 1972. *Sociolinguistic patterns*. Oxford: Basil Blackwell.

Labov, William. 1981. Resolving the neogrammarian controversy. *Language* 57. 267–308.

Labov, William. 1994. *Principles of linguistic change, vol. 1: Internal factors*. Oxford: Blackwell.

Labov, William. 2001. *Principles of linguistic change, vol. 2: Social factors*. Oxford: Blackwell.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 133–150. Amsterdam & New York: Rodopi.

Lefebvre, Claire. 2015. *Relabeling in language genesis*. Oxford: Oxford University Press.

Lehmann, Thomas. 1993. *A grammar of modern Tamil*. 2nd edn. Pondicherry: Pondicherry Institute of Linguistics & Culture.

Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, variation and usage*. 1st publ. Cambridge: Cambridge University Press.

Leung, Yan-kit Ingrid. 2003. Failed features versus full transfer full access in the acquisition of a third language: Evidence from tense and agreement. In *Proceedings of the 6th Generative Approaches to a Second Language Acquisition Conference (GASLA 2002)*, 199–207. Somerville, MA: Cascadilla Press.

Li, Yen-Hui Audrey. 1986. Abstract case in Chinese. *Dissertation Abstracts International* 46(9).

Li, Yen-Hui Audrey. 1990. *Order and constituency in Mandarin Chinese*. Dordrecht: Kluwer.

Lim, Lisa (ed.). 2004. *Singapore English: A grammatical description*. Amsterdam & Philadelphia: John Benjamins.

Lim, Lisa. 2007. Mergers and acquisitions: On the ages and origins of Singapore English particles. *World Englishes: Journal of English as an International and Intranational Language* 26(4). 446–473.

Lim, Lisa. 2009. *The Grammar of Spoken Singapore English Corpus (GSSEC): Ground rules & conventions*. https://english.hku.hk/staff/lisa_lim/GSSEC-GroundRules-2009.doc.

Lim, Lisa & Joseph A. Foley. 2004. English in Singapore and Singapore English: Background and methodology. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 1–18. Amsterdam: John Benjamins.

Lim, Lisa & Nikolas Gisborne. 2009. The typology of Asian Englishes: Setting the agenda. *English World-Wide* 30(2). 123–132.

Liu, Jing, Evie Tindall & Deanna Nisbet. 2006. Chinese learners and English plural forms. *Linguistics Journal* 1(3). 127–147.

Lombardi, Linda. 1991. *Laryngeal features and laryngeal neutralization*. Amherst, MA: University of Massachusetts. PhD thesis.

Low, Ee Ling. 2014. Research on English in Singapore. *World Englishes* 33(4). 439–457.

Lüdeling, Anke, Stefan Evert & Marco Baroni. 2007. Using web data for linguistic purposes. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 7–24. Amsterdam & New York: Rodopi.

Mair, Christian. 2013. 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.

Mair, Christian. 2018. Stabilising domains of English-language use in Germany. In Sandra C. Deshors (ed.), *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, 45–76. Amsterdam: John Benjamins.

Mair, Christian & Marianne Hundt. 2000. Introduction: Corpus linguistics and linguistic theory. In Christian Mair & Marianne Hundt (eds.), *Corpus linguistics and linguistic theory: Papers from the twentieth International Conference on English Language Research on Computerised Corpora (ICAME 20), Freiburg im Breisgau 1999*. Amsterdam & Atlanta, GA: Rodopi.

Matras, Yaron. 2009. *Language contact*. Cambridge: Cambridge University Press.

Matthews, Stephen J. & Virginia Yip. 1994. *Cantonese*. London: Routledge.

Max Planck Institute for Psycholinguistics. 2001. *WebCelex*. http://celex.mpi.nl/ (8 February, 2016).

McArthur, Tom. 1998. *The English languages*. Cambridge: Cambridge University Press.

McArthur, Tom. 2002. *The Oxford guide to world English*. Oxford: Oxford University Press.

McCarthy, John J. & Alan S. Prince. 1994. The emergence of the unmarked: Optimality in prosodic morphology. In Merce González (ed.), *Proceedings of the North East Linguistic Society 24*, 333–379. Amherst, MA: GLSA.

McWhorter, John H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2). 125–166.

Medin, Douglas L. & Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85(3). 207–238.

Meierkord, Christiane. 2012. *Interactions across Englishes: Linguistic choices in local and international contact situations*. Cambridge: Cambridge University Press.

Mesthrie, Rajend (ed.). 2008. *Varieties of English, 4: Africa, South and Southeast Asia*. Berlin: Mouton de Gruyter.

Mesthrie, Rajend. 2012. Regional Profile: Asia. In Bernd Kortmann (ed.), *The Mouton World Atlas of Variation in English*, 784–805. Berlin: De Gruyter Mouton.

Miestamo, Matti. 2009. Implicational hierarchies and grammatical complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 80–97. Oxford: Oxford University Press.

Milroy, Lesley. 1980. *Language and social networks*. Oxford: Blackwell.

Milroy, Lesley & Matthew Gordon. 2003. *Sociolinguistics: Method and interpretation*. 1st publ. Malden, MA: Blackwell.

Mufwene, Salikoko S. 2001. *The ecology of language evolution*. Cambridge: Cambridge University Press.

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 & Marianne Hundt (eds.). 2011. *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*. Amsterdam: John Benjamins.

Muysken, Pieter. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge: Cambridge University Press.

Nelson, Gerald. 2002. *Markup manual for spoken texts*. http://www.ice-corpora.uzh.ch/dam/jcr:72c70d5a-8da8-496f-b8dc-5fb66986c87c/spoken.pdf (9 September, 2018).

Nesselhauf, Nadja. 2009. Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide* 30(1). 1–26.

Nihalani, Paroo, Ray K. Tongue & Priya Hosali. 1979. *Indian and British English: A handbook of usage and pronunciation*. New Dehli: Oxford University Press.

Nihalani, Paroo, Ray K. Tongue, Priya Hosali & Jonathan Crowther. 2004. *Indian and British English: A handbook of usage and pronunciation*. 2nd edn. New Delhi: Oxford University Press.

Nikolaeva, Irina. 2007. *Finiteness: Theoretical and empirical foundations*. Oxford: Oxford University Press.

Odlin, Terence. 1989. *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Office of the Registrar General and Census Commissioner, India. 2018a. *Census of India: Statement 4. Scheduled languages in descending order of speakers' strength—2011*. http://www.censusindia.gov.in/2011Census/Language-2011/Statement-4.pdf (16 September, 2018).

Office of the Registrar General and Census Commissioner, India. 2018b. *Census of India: Statement 5. Comparative speakers' strength of scheduled languages—1971, 1981, 1991, 2001 and 2011*. http://www.censusindia.gov.in/2011Census/Language-2011/Statement-5.pdf (16 September, 2018).

Onysko, Alexander. 2016. Modeling world Englishes from the perspective of language contact. *World Englishes* 35(2). 196–220.

Ooi, Vincent B. Y. 2001a. Ethnic group varieties of Singapore English: Melody or harmony? In Vincent B. Y. Ooi (ed.), *Evolving identities: The English language in Singapore and Malaysia*, 53–68. Singapore: Times Academic Press.

Ooi, Vincent B. Y. (ed.). 2001b. *Evolving identities: The English language in Singapore and Malaysia*. Singapore: Times Academic Press.

Oxford English Dictionary, 2nd edn. CD-Rom. 1999. Oxford: Oxford University Press.

Pakir, Anne. 1991. The range and depth of English-knowing bilinguals in Singapore. *World Englishes* 10(2). 167–179.

Pan, Lynn (ed.). 1999. *The encyclopedia of the Chinese overseas*. Singapore: Landmark Books.

Peng, Long & Jane Setter. 2000. The emergence of systematicity in the English pronunciations of two Cantonese-speaking adults in Hong Kong. *English World-Wide* 21(1). 81–108.

Pfaff, Carol W. (ed.). 1987. *First and second language acquisition processes*. Cambridge, MA: Newbury House.

Phillips, Betty S. 1984. Word frequency and the actuation of sound change. *Language* 60(2). 320–342.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–158. Amsterdam: John Benjamins.

Platt, John T. & Heidi Weber. 1980. *English in Singapore and Malaysia. Status, features, functions*. New York: Oxford University Press.

Platt, John T., Heidi Weber & Mian Lian Ho. 1984. *The New Englishes*. London: Routledge & Kegan Paul.

Posner, Michael I. & Steven W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77(3). 353–363.

Pride, John B. 1982. *New Englishes*. Rowley, MA: Newbury House.

Prince, Alan S. & Paul Smolensky. 1993. *Optimality theory: Constraint interaction in generative grammar*. Report no. RuCCS-TR-2. New Brunswick, NJ: Rutgers University Center for Cognitive Science.

Prolific. 2016. *Participant pool demographics: Country of birth*. https://www.prolific.ac/demographics?metric=54ac6ea9fdf99b2204feb895 (13 March, 2017).

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Ramson, William S. 1966. *Australian English: An historical study of the vocabulary 1788–1898*. Canberra: Australian National University Press.

Randall, Mick. 1997. Orthographic knowledge, phonological awareness and the teaching of English: An analysis of word dictation errors in English of Malaysian secondary school pupils. *RELC Journal* 28(2). 1–21.

Randall, Mick. August 2003. Dictation errors and what they can tell us about lexical representations; pedagogical implications. Paper presentation. 3rd International Literacy Conference, Literacy: Bridging Past, Present and Future 15-17 August. Penang, Malaysia.

Rosso, Mark A. & Stephanie W. Haas. 2010. Identification of web genres by user warrant. In Alexander Mehler, Serge Sharoff & Marina Santini (eds.), *Genres on the web: computational models and empirical studies*, 47–68. New York: Springer.

RStudio. 2016. *Home*. RStudio. https://www.rstudio.com/ (1 May, 2017).

Rüdiger, Sofia. 2017. *Characterizing the Spoken Korean English repertoire: Morphosyntactic patterns of Korean(ized) English*. Universität Bayreuth. PhD thesis.

Sahgal, Anju & Rama K. Agnihotri. 1985. Is Indian English retroflexed and r-full. *Indian Journal of Applied Linguistics* 11. 97–109.

Sailaja, Pingali. 2009. *Indian English*. Edinburgh: Edinburgh University Press.

Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language complexity as an evolving variable*. Oxford: Oxford University Press.

Sand, Andrea. 2005. *Angloversals? Shared morphosyntactic features in contact varieties of English*. University of Freiburg Habilitationsschrift.

Saravanan, Vanithamani. 1989. *Variation in Singapore Tamil English*. Monash University. Unpublished doctoral dissertation.

de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Paris: Payot.

Saw, Swee Hock. 1999. *The population of Singapore*. Singapore: Institute of Southeast Asian Studies.

Schiffman, Harold F. 1983. *A reference grammar of spoken Kannada*. Seattle: University of Washington Press.

Schneider, Edgar W. (ed.). 1997. *Englishes around the world. Vol. 1: General studies, British Isles, North America. Vol. 2: Caribbean, Africa, Asia, Australasia. Studies in honour of Manfred Görlach. Varieties of English around the world G18, G19*. Amsterdam: John Benjamins.

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, NY: Cambridge University Press.

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. 2014. New reflections on the evolutionary dynamics of world Englishes. *World Englishes* 33(1). 9–32.

Schröter, Verena. 2017. *"Where got such thing one?" - A multivariate analysis of null subjects in Asian Englishes*. Albert-Ludwigs-Universität Freiburg. PhD thesis.

de Schryver, Gilles-Maurice. 2002. Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies* 11(2). 266–282.

Schuchardt, Hugo. 1885. *Über die Lautgesetze: gegen die Junggrammatiker*. Berlin: Oppenheim.

Schütze, Carson T. & Jon Sprouse. 2013. Judgment data. In Robert J. Podesva & Devyani Sharma (eds.), *Research Methods in Linguistics*, 27–50. Cambridge: Cambridge University Press.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and change*. Amsterdam & Philadelphia: John Benjamins.

Setter, Jane, Cathy S. P. Wong & Brian Hok-Shing Chan. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.

Sharma, Devyani. 2005. Language transfer and discourse universals in Indian English article use. *Studies in Second Language Acquisition* 27(4). 535–566.

Sharma, Devyani. 2009. Typological diversity in New Englishes. *English World-Wide* 30(2). 170–195.

Sharma, Devyani & Ashwini Deo. 2009. Contact-based aspectual restructuring: A critique of the Aspect Hypothesis. *Queen Mary Occasional Papers Advancing Linguistics* (OPALS) series.

Sharoff, Serge. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni & Silvia Bernardini (eds.), *WaCky! Working papers on the web as corpus*, 63–98. Bologna: Gedit.

Sheng, He Ji. 2007. *Grammatical features of Singapore Colloquial English: A corpus-based variation study*. National University of Singapore. PhD thesis.

Siegel, Jeff. 1999. Transfer constraints and substrate influence in Melanesian Pidgin. *Journal of Pidgin and Creole Languages* 14(1). 1–44.

Siemund, Peter. 2013. *Varieties of English: A typological approach*. Cambridge: Cambridge University Press.

Silver, Rita E., Lubna Alsagoff & Christine C. M. Goh (eds.). 2009. *Language learning in new English contexts: Studies of acquisition and development*. London: Continuum.

Sinclair, John M. 2001. *Collins Cobuild English dictionary for advanced learners*. 3rd edn. London: HarperCollins.

Slobin, Dan I. 1973. Cognitive prerequisites for the development of grammar. In Charles A. Ferguson & Dan I. Slobin (eds.), *Studies of Child Language Development*. 175–208. New York: Holt.

Slobin, Dan I. 1977. Language change in childhood and in history. In John Theodore Macnamara (ed.), *Language learning and thought*, 185–214. New York: Academic Press.

So, Daniel. 1992. Language-based bifurcation of secondary education in Hong Kong: Past, present and future. In Kang Kwong Kapathy Luke (ed.), *Into the twenty first century: Issues of language and education in Hong Kong*, 69–95. Hong Kong: Linguistic Society of Hong Kong.

Spencer, John (ed.). 1971. *The English language in West Africa*. London: Longman.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134. 219–248.

Sridhar, Kamal K. & Shikaripur N. Sridhar. 1986. Bridging the paradigm gap: Second-language acquisition theory and indigenized varieties of English. *World Englishes* 5(1). 3–14.

Sridhar, Shikaripur N. 1992. The ecology of bilingual competence: Language interaction in the syntax of indigenized varieties of English. *World Englishes* 11(2). 141–150.

Staum Casasanto, Laura, Philip Hofmeister & Ivan A. Sag. 2010. Understanding acceptability judgments: Distinguishing the effects of grammar and processing on acceptability judgments. In Stellan Ohlsson & Richard Catrambone (eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Stemberger, Joseph P. & Brian MacWhinney. 1988. Are inflected forms stored in the lexicon? In Michael Hammond & Michael Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics*, 101–116. San Diego, CA: Academic Press.

Steriade, Donca. 1982. *Greek prosodies and the nature of syllabification*. Published 1990. Cambridge, MA: Massachusetts Institute of Technology. PhD thesis.

Steriade, Donca. 2001. *The phonology of perceptibility effects: The P-map and its consequences for constraint organization*. UCLA. Unpublished Manuscript.

von Sutterheim, Christiane & Wolfgang Klein. 1987. A concept-oriented approach to second language acquisition studies. In Carol W. Pfaff (ed.), *First and second language acquisition processes*. Cambridge, MA: Newbury House.

de Swaan, Abram. 2002. *The World Language System: A political sociology and political economy of language*. Cambridge: Polity.

de Swaan, Abram. 2010. Language systems. In Nikolas Coupland (ed.), *The handbook of language and globalization*, 56–76. Malden, MA: Wiley-Blackwell.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. Between simplification and complexification: Non-standard varieties of English around the world. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*. Oxford: Oxford University Press.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2012. Introduction: Linguistic complexity. In Bernd Kortmann & Benedikt Szmrecsanyi (eds.), *Linguistic complexity: Second language acquisition, indigenization, contact*, 6–34. Berlin: De Gruyter.

Tay, Mary W. J. 1979. The uses, users and features of English in Singapore. In Jack C. Richards (ed.), *New varieties of English*, 91–111. Singapore: SEAMCO Regional Language Centre.

The ICE Project. 2016. *International Corpus of English (ICE)*. International Corpus of English. http://ice-corpora.net/ice/ (16 September, 2018).

The R Foundation. 2017. *R: The R Project for Statistical Computing*. https://www.r-project.org/ (1 May, 2017).

Tickoo, Asha. 2005. The selective marking of past tense: Insights from Indian learners of English. *International Journal of Applied Linguistics* 15(3). 364–378.

Tongue, Ray K. 1974. *The English of Singapore and Malaysia*. Singapore: Eastern University Press.

Trask, Robert L. 2000. *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.

Trudgill, Peter. 2001. Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5(2). 371–374.

Trudgill, Peter. 2009. Vernacular universals and the sociolinguistic typology of English dialects. In Markku Filppula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular universals and language contacts: Evidence from varieties of English and beyond*, 304–322. London: Routledge.

Trudgill, Peter. 2010. Contact and sociolinguistic typology. In Raymond Hickey (ed.), *The handbook of language contact*, 299–319. Oxford: Wiley-Blackwell.

Trudgill, Peter & Jean Hannah (eds.). 1982. *International English: A guide to varieties of Standard English*. London: Arnold.

Tsui, Amy B. M. & David Bunton. 2000. The discourse and attitudes of English language teachers in Hong Kong. In Kingsley Bolton (ed.), *Hong Kong English: Autonomy and creativity*, 287–303. Special Issue of World Englishes 19(3). Published as a book 2002. Hong Kong: Hong Kong University Press.

Turnbull, Constance M. 1977. *A history of Singapore: 1819–1975*. Singapore: Oxford University Press.

Turner, George W. 1966. *The English language in Australia and New Zealand*. London: Longman.

Universitätsbibliothek Freiburg. 2018. *Freiburg Corpus of English Dialects (FRED) - Interactive Database*. https://fred.ub.uni-freiburg.de/ (20 July, 2018).

Vaish, Viniti. 2008. *Biliteracy and globalization: English language education in India*. Clevedon: Multilingual Matters.

Van Rooy, Bertus. 2011. A principled distinction between error and conventionalized innovation in African Englishes. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 189–208. Amsterdam: John Benjamins.

Véronique, Daniel. 1987. Reference to past events and actions in narratives in L2: Insights from North African learners' French. In Carol W. Pfaff (ed.), *First and second language acquisition processes*. Cambridge, MA: Newbury House.

Walker, Anthony R. 2004. South Asians in Malaysia and Singapore. In Melvin Ember, Carol R. Ember & Ian Skoggard (eds.), *Encyclopedia of diasporas: Immigrant and refugee cultures around the world*, 1105–1119. New York, NY: Kluwer Academic.

Wang, William S.-Y. 1969. Competing changes as a cause of residue. *Language* 45(1). 9–25.

Wang, William S.-Y. (ed.). 1977. *The lexicon in phonological change*. The Hague: De Gruyter Mouton.

Wang, William S.-Y. & Chin-Chuan Cheng. 1977. Implementation of phonological change: The Shaungfeng Chinese case. In William S.-Y. Wang (ed.), *The lexicon in phonological change*, 86–100. The Hague: De Gruyter Mouton.

Warren, Tessa, Sarah J. White & Erik D. Reichle. 2009. Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition* 111(1). 132–137.

Wee, Lionel & Umberto Ansaldo. 2004. Nouns and noun phrases. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 57–74. Amsterdam & Philadelphia: John Benjamins.

Wells, John C. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Williams, Jessica. 1987. Non-native varieties of English: A special case of language acquisition. *English World-Wide: A Journal of Varieties of English* 8(2). 161–199.

Wilson, Sheila & Rajend Mesthrie. 2004. St. Helena English: morphology and syntax. In Bernd Kortmann & Edgar W. Schneider (eds.), *A handbook of varieties of English. A multimedia reference tool. Volume 2: Morphology and syntax*, 1006–1015. Berlin & New York: Mouton de Gruyter.

Wiltshire, Caroline R. 2013. Emergence of the unmarked in Indian Englishes with different substrates. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of World Englishes*, 599–620. Oxford: Oxford University Press.

Wiltshire, Caroline R. 2014. New Englishes and the emergence of the unmarked. In Eugene Green & Charles F. Meyer (eds.), *The variability of current world Englishes*, 13–38. Berlin: De Gruyter Mouton.

Winford, Donald. 2003. *An introduction to contact linguistics*. 1st publ. Oxford: Blackwell Publishing.

Wong, May. 2017. *Hong Kong English: Exploring lexicogrammar and discourse from a corpus-linguistic perspective*. London: Palgrave Pivot.

Yadurajan, Katticenahalli S. 2001. *Current English: A guide for the user of English in India*. 1st publ. New Delhi: Oxford University Press.

Yap, Dennis S. B. 2006. *Errors in tense and aspect in the compositions of secondary school pupils*. Singapore: National Institute of Education, Nanyang Technological University. Master's thesis.

Yeo, Josephine N. P. & David Deterding. 2003. Influences of Chinese and Malay on the written English of secondary students in Singapore. In David Deterding, Ee Ling Low & Adam Brown (eds.), *English in Singapore: Research on grammar*, 77–84. Singapore: McGraw-Hill.

Yip, Virginia. 2004. *Errors in past tense marking: A study of Primary 5 students in Singapore*. Singapore: National Institute of Education, Nanyang Technological University Honours Academic Exercise.

Ziegeler, Debra. 2015. *Converging grammars: Constructions in Singapore English*. Berlin & Boston: De Gruyter Mouton.

Zipf, George K. 1935. *The psychobiology of language*. Boston, MA: Houghton Mifflin.

German summary

Dieses Buch beschäftigt sich mit der Frage, inwiefern Gebrauchsfrequenzen Simplifizierungsprozesse in asiatischen Varietäten des Englischen erklären können, und leistet somit einen Beitrag zur Sprachkontaktforschung, insbesondere der Erforschung von kontaktinduziertem morphosyntaktischen Sprachwandel. Die betrachteten Varietäten sind das Englische in Hong Kong (HKE), Indien (IndE) und Singapur (SgE). Das Britische Englisch (BrE) und das Amerikanische Englisch (AmE) dienen als Kontrollvarietäten.

Die Grundannahme ist, dass in Varietäten, in denen sich ein Simplifzierungsprozess durchgesetzt hat, frequente Formen stärker von der Simplifizierung betroffen sind als weniger frequente Formen. Umgekehrt wird vermutet, dass in Varietäten, in denen sich ein Simplifizierungsprozess nicht durchgesetzt hat, frequente Formen weniger stark von der Simplifizierung betroffen sind als weniger frequente Formen. Diesen Annahmen liegt zugrunde, dass häufig verwendete Formen besonders stark im Gedächtnis verankert sind (vgl. Hockett 1958: 180–181, in Bybee 1985: 119; Bybee & Thompson 2007: 271). Sind die betrachteten Simplifizierungsprozesse in einer Varietät etabliert, so wird erwartet, dass sich simplifizierte Formen, die häufig verwendet werden, auch besonders rasch durchsetzen. Sind sie nicht etabliert und Simplifizierung taucht nur sporadisch auf, so wird vermutet, dass insbesondere wenig frequente Formen vereinfacht werden.

Bei den betrachteten Simplifizierungsprozessen handelt es sich um den Wegfall der Vergangenheitsmarkierung bei Verben und den Wegfall der Pluralmarkierung bei Nomen, zwei Reduktionsphänomene, die mit struktureller Vereinfachung einhergehen. Außerdem werden zwei Beispiele für Regularisierung untersucht, nämlich die Regularisierung der Vergangenheitsmarkierung irregulärer Verben und der Gebrauch von nicht zählbaren Nomen als zählbare Nomen. Regularisierung führt nicht (zwingend) zu struktureller Vereinfachung, sondern zu mehr Transparenz im bestehenden System, zum Beispiel indem irreguläre Verben ihre Vergangenheitsform durch Anhängen des *–ed* Suffixes an den Verbstamm bilden und sich so der Mehrheit der Verben, die ihre Vergangenheitsform regulär bilden, anpassen. Reduktion und

Regularisierung sind Simplifizierungsprozesse, die typischerweise in Sprachkontaktsituationen auftreten (vgl. Trudgill 2001: 372–373).

Die betrachteten Simplifizierungsprozesse sind insbesondere für SgE in der Literatur belegt (etwa Gut 2009b; Ziegeler 2015: 182–183), jedoch gibt es verhältnismäßig wenige detaillierte oder gar empirische Studien zu den Phänomenen. Low (2014) spricht gar von der dringenden Notwendigkeit einer empirischen Validierung theoriebasierter Erkenntnisse zu SgE („an urgent need for empirical validation"; 454) und schlägt Vergleiche zu anderen Varietäten des Englischen vor. Auch in Bezug auf HKE, welches weit weniger gut erforscht ist als SgE, werden die Phänomene zwar vereinzelt genannt (etwa Setter u. a. 2010: 45–49; Budge 1989: 41), jedoch kaum empirisch untermauert. Laut eWAVE (*electronic World Atlas of Varieties of English*; Kortmann & Lunkenheimer 2013), einer interaktiven Plattform, die für 50 Varietäten des Englischen sowie für 26 auf Englisch basierende Pidgin- und Kreolsprachen Informationen zum Auftreten spontansprachlicher morphosyntaktischer Merkmale bietet, sind die betrachteten Phänomene typischerweise in HKE und in CSE (Colloquial SgE) zu finden, jedoch nur selten in IndE.

Die Varietätenkonstellation aus HKE, IndE und SgE wurde aus Gründen des Substrateinflusses und des Institutionalisierungsgrades der Varietäten gewählt. HKE und SgE haben Mandarin und Kantonesisch als als isolierende Substratsprachen gemein, während IndE insbesondere auf agglutinierenden (Tamil) und flektierenden (Hindi) Substratsprachen beruht (vgl. L. Lim & Gisborne 2009: 126). Unterschiede zwischen HKE und SgE auf der einen Seite und IndE auf der anderen Seite, wie sie in eWAVE auftauchen, sind also möglicherweise zum Teil mit unterschiedlichem Substrateinfluss erklärbar. Ein Vergleich zwischen HKE und SgE bietet sich an, da beide Varietäten eine ähnliche Kontaktökologie besitzen, jedoch trotz ähnlicher Kolonialgeschichte sehr unterschiedliche Institutionalisierungsgrade aufweisen (vgl. Schneider 2007: 138–139, 156–160). SgE hat sich in Singapur zur Alltagssprache mit identitätsstiftender Umgangssprache entwickelt, während in Hong Kong der Gebrauch des Englischen weitestgehend auf den administrativen, schulischen und beruflichen Kontext begrenzt ist.

Die Gebrauchsfrequenzen in den betrachteten Varietäten werden anhand verfügbarer und weit genutzter Korpora im Bereich *World Englishes*, nämlich dem *International Corpus of English* (ICE; Greenbaum 1996) und dem *Corpus of Global Web-Based English* (GloWbE; Davies 2013), approximiert. Was ICE betrifft, werden nur die gesprochensprachlichen Teile des Korpus betrachtet (600.000 Wörter pro

Varietät). GloWbE ist ein Internetkorpus und umfasst etwa 42 Millionen Wörter für HKE und SgE sowie etwa 96 Millionen Wörter für IndE. Als Frequenzmaß werden Lemmatokenfrequenzen hergenommen, da sich Worttokenfrequenzen für simplifizierte und nicht-simplifizierte Formen unterscheiden und den Analysen somit unterschiedliche Frequenzwerte zugrunde gelegt werden müssten. In Bezug auf den Wegfall der Vergangenheitsmarkierung bei Nomen werden die Wegfallraten zudem auf der Basis von Typfrequenzen untersucht, die wie die Lemmatokenfrequenzen anhand von GloWbE approximiert werden.

Das Vorkommen der relevanten Phänomene wurde in den drei Zielvarietäten anhand von ICE und GloWbE ermittelt. Die Wegfallraten der Vergangenheits- und Pluralmarkierung sowie die Regularisierungsraten wurden hierzu anhand randomisierter Stichproben regulärer und irregulärer Verben sowie Nomen mit regulärer Pluralbildung erhoben. Für ICE zeigt sich, dass die Phänomene insbesondere in HKE und in geringerem Ausmaß auch in SgE zwar auftreten, jedoch zu selten sind, als dass man von etablierten und für die Varietäten typischen Merkmalen sprechen könnte. Frequenzeffekte dahingehend, dass bei nicht etablierten Simplifizierungsprozessen frequente Formen weniger stark von der Simplifizierung betroffen sind als weniger frequente Formen (siehe oben), lassen sich aufgrund der generell niedrigen Wegfall- und Regularisierungsraten schwer ausmachen. In HKE sind jedoch für einige frequente Verben und Nomen vergleichsweise hohe Wegfallraten feststellbar. In GloWbE kommen die untersuchten Phänomene kaum vor, was zeigt, dass es sich um rein gesprochensprachliche Merkmale handelt, die in geschriebener Form (hier: geschriebene Sprache im Internet) fast keine Rolle spielen.

In Bezug auf den Einfluss von Substrateinfluss und Institutionalisierungsgrad lässt sich Folgendes festhalten: Während die sehr geringen Regularisierungsraten es nicht erlauben, von klarem Substrateinfluss zu sprechen, zeigt sich der Substrateinfluss beim Wegfall der Vergangenheits- und der Pluralmarkierung deutlicher. Besonders im SgE ist der Wegfall der Vergangenheitsmarkierung auf die Reduktion von Konsonantenclustern zurückzuführen. Dies ist auch im HKE der Fall, jedoch ist hier der Unterschied im Wegfall der Flexionsmarkierung zwischen Verben, die auf Konsonanten enden (etwa *call*), und solchen, die auf Vokale enden (etwa *play*), weit weniger stark ausgeprägt. Auch ist insbesondere im HKE ein Wegfall der Pluralmarkierung bei Nomen dann zu beobachten, wenn ein Zahlwort (z.B. *many*) vorausgeht. Das Kantonesische markiert Plural entsprechend, was für Substrateinfluss spricht. Was den Einfluss des Institutionalisierungsgrades angeht, lässt sich festhalten, dass der

Wegfall der Vergangenheits- und Pluralmarkierung im HKE vergleichsweise willkürlich zu sein scheint, was aufgrund des geringen Institutionalisierungsgrades dieser Varietät erwartet wurde. Im Gegensatz dazu folgt SgE, welches weit stärker etabliert ist, klareren Mustern.

In einem letzten Schritt wird ein webbasiertes Perzeptionsexperiment bestehend aus einer *self-paced reading* Aufgabe und einer *acceptability judgment* Aufgabe präsentiert, das den Einfluss vorangehender Zeitadverbialen mit Vergangenheitsbezug und Zahlwörter mit Pluralbezug auf die Perzeption von Verben und Nomen ohne Vergangenheits- bzw. Pluralmarkierung untersucht. Sprecher des HKE, des IndE und des SgE bilden die Zielgruppen, Sprecher des BrE und des AmE die Kontrollgruppe. Die Grundannahme ist, dass Sprecher des HKE und des SgE Verben und Nomen ohne Flexionsmarkierung vergleichsweise schnell lesen (*self-paced reading*) und vergleichsweise positiv bewerten (*acceptability judgment*). Für diese Varietäten ist bekannt, dass der Vergangenheits- bzw. Pluralbezug oft anhand von entsprechenden Zeitadverbialen bzw. Zahlwörtern erfolgt, was die Flexionsmarkierung für Verben und Nomen streng genommen redundant macht (etwa Bao 1998: 163; Wee & Ansaldo 2004: 64; Setter u. a. 2010: 45–49). Ein Vergleich beider Sprechergruppen in den zwei verwendeten Perzeptionsaufgaben bietet sich an, da eine gewisse Vertrautheit mit unmarkierten Verben und Nomen in beiden Gruppen wahrscheinlich ist, jedoch angenommen wird, dass Sprecher des SgE die nicht dem Standard entsprechenden Formen aufgrund des hohen Institutionalisierungsgrades ihrer Varietät besonders akzeptieren.

Das Perzeptionsexperiment zeigt, dass Sprecher aller drei Kontaktvarietäten Verben ohne Flexionsmarkierung und die direkt im Satz folgenden Worte schneller lesen und besser bewerten als die Kontrollgruppe, unabhängig davon, ob eine Zeitadverbiale vorangeht oder nicht. Nomen ohne Flexionsmarkierung und die ihnen folgenden Worte werden von den Zielgruppen vergleichsweise schnell gelesen und vergleichsweise gut bewertet, wenn kein Zahlwort vorausgeht. Letzteres mag daran liegen, dass in Sätzen wie *Ben told them detail about...* statt der fehlenden Pluralmarkierung auch der Artikel fehlen könnte, ein Phänomen, welches insbesondere für SgE bekannt ist (etwa Siemund 2013: 99). Die Ergebnisse zeigen also, dass unabhängig vom Vorkommen einer Zeitadverbiale oder eines Zahlwortes unmarkierte Verben und Nomen von Sprechern der Zielgruppen verhältnismäßig schnell gelesen und gut bewertet werden. Ein nennenswerter Frequenzeffekt ist dahingehend zu beobachten, dass sprechergruppenübergreifend unmarkierte Verben und Nomen (und die ihnen

folgenden Wörter), die häufig verwendet werden, besonders lange Lesezeiten aufweisen.

Zusammenfassend lässt sich festhalten, dass die relevanten Simplifizierungsprozesse in der betrachteten Stichprobe von Verben und Nomen insbesondere in HKE und SgE zwar auftreten, die häufige Nennung der Phänomene in der Literatur (besonders der zu SgE) aber wohl eher auf ihre Salienz (im Sinne von Auffälligkeit) als auf ihre Frequenz zurückzuführen ist. In der Tat zeugen die vergleichsweise schnellen Lesezeiten und guten Bewertung der unmarkierten Formen der Singapurer Teilnehmer im Experiment von einer gewissen Vertrautheit dieser Sprechergruppe mit dem Wegfall der Vergangenheitsmarkierung bei Verben und dem Wegfall der Pluralmarkierung bei Nomen.

Das Buch greift den Einfluss von Substrateinfluss und Institutionalisierungsgrad auf die Entwicklung von Kontaktphänomenen auf und erweitert die Diskussion um einen frequenzbasierten Ansatz. Die Entwicklung hin zu neueren und größeren Korpora (etwa GloWbE) bietet die Möglichkeit, den frequenzbasierten Ansatz für gut und weniger gut beschriebene Phänomene zu testen und ihn auch auf seltenere Merkmale (z.B. die betrachteten Regularisierungsprozesse) auszuweiten. Ein Ziel des Buches ist es außerdem, die Verknüpfung von Korpusstudien mit experimentellen Studien zu motivieren. Letztere erlauben es, aus Korpusanalysen entstandene Vermutungen unter kontrollierten Bedingungen zu testen, was im vorliegenden Fall um die Verbindung von Produktionsstudien (Korpusanalysen) mit einer Perzeptionsstudie (Experiment) ergänzt wird.

This book presents a corpus- and usage-based approach to morphological simplification in three Asian Englishes (Hong Kong English, Indian English, and Singapore English). The features of interest are omission of verbal past tense marking, omission of nominal plural marking, regularization of irregular verbs, and the use of uncountable nouns as countable nouns. Usage frequency, substratum transfer, and institutionalization are investigated as potential determinants of simplification. Of particular interest is the question in how far substratum transfer and institutionalization constrain frequency effects.

To account for the production of the features of interest, the *International Corpus of English* (ICE) and the *Corpus of Global Web-Based English* (GloWbE) are used. A perception experiment accompanying the corpus studies on verbal past tense and nominal plural marking provides additional insights into the perception of morphological simplification in the Asian Englishes considered.

The impact of usage frequencies on feature development in World Englishes has received little attention so far. In pursuing a systematic quantitative approach to omission and regularization patterns, the present study aims to establish usage-based thinking within World Englishes research. The distinction between regular and irregular forms is a prime example of frequency-driven language development, which is why studying omission and regularization marks an ideal starting point for this endeavor.

Laura A. M. Terassa obtained a bachelor's degree in English Studies and Economics from the University of Erlangen-Nuremberg (Germany) in 2010 and a master's degree in English Language and Linguistics from the University of Freiburg (Germany) in 2013. From 2014 to 2017, she was a member of the doctoral research training group "DFG GRK 1624: Frequency effects in language" at the University of Freiburg. This book is a revised version of her dissertation, which she defended in February 2018.

Die Publikationsreihe NIHIN – New Ideas in Human Interaction – entstand 2010 und ist ein Kooperationsprojekt zwischen der Hermann Paul School of Linguistics (HPSL) und der Universitätsbibliothek Freiburg (UB).

NIHIN bietet eine moderne, frei zugängliche Plattform für wissenschaftliche Essays erfahrener WissenschaftlerInnen sowie Prädikatsdissertationen, Textsammlungen zum Thema Sprache in der Interaktion und multimodale Sprachkorpora.

ISBN 978-3-928969-74-1

eucor The European Campus

Universität Basel

UNI FREIBURG