

Accepted Manuscript. Published at:

Tamaredo, Ivan. (2020). *Complexity, efficiency, and language contact: Pronoun omission in World Englishes*. (Linguistic Insights 270). Bern: Peter Lang. DOI: <https://doi.org/10.3726/b16943>

TABLE OF CONTENTS

List of figures	v
List of tables	vii
Acknowledgements	x
List of abbreviations	xi
1 Introduction	1
1.1 Communicative efficiency and language complexity	1
1.2 Aims and research questions	5
1.3 Structure of the book	8
2 Language complexity	12
2.1 Historical background	13
2.2 Sources of complexity variance.....	18
2.2.1 Language contact.....	20
2.2.2 Acquisition versus use	24
2.2.3 A typology of contact situations.....	28
2.2.4 Additional social determinants of complexity.....	30

2.3 Measuring complexity	31
2.3.1 Kusters (2003, 2008).....	31
2.3.2 Miestamo (2006a, 2008)	34
2.3.3 Hawkins (1994, 2004, 2009, 2014).....	35
2.3.4 Comparison between the three metrics	39
2.3.5 Some important distinctions	43
2.4 Complexity and varieties of English.....	47
3 Pronoun omission.....	55
3.1 Generative approaches	57
3.2 Cognitive approaches.....	62
3.2.1 Agreement and context	63
3.2.2 Accessibility theory	67
3.2.3 Further factors.....	70
3.3 Pronoun omission in English	74
3.3.1 Pronoun omission in the history of British English.....	78
3.3.2 Constraints in Present-day English	80
3.4 The complexity and efficiency of pronoun omission	87
4 A cross-varietal study of pronoun omission in English	96
4.1 Simplification and substrate effects	99
4.2 Data and methodology	103

4.3 Results	110
4.3.1 Global indexes: Pronoun omission attestation and pervasiveness	111
4.3.2 Attestation and pervasiveness of individual features	120
4.4 Simplification and substrate effects revisited.....	128
5 Complexity variance in English: Pronoun omission and language contact	132
5.1 Aims and purpose of the study	133
5.2 Data and methodology.....	137
5.2.1 Corpus and data retrieval process	137
5.2.2 Varieties selected	142
5.2.3 Variables included in the analysis	145
5.2.3.1 Language-external constraints.....	145
5.2.3.2 Language-internal constraints	146
5.2.3.3 Interim summary	159
5.2.4 Statistical analysis	161
5.2.4.1 Regression modelling and structural complexity	162
5.2.4.2 Random forests and system complexity	164
5.3 Frequency of omitted and overt pronouns per variety.....	167
5.3.1 Results	167
5.3.2 Discussion	186
5.4 Multivariate analysis of the data.....	189

5.4.1 Structural complexity.....	189
5.4.1.1 Results.....	189
5.4.1.2 Discussion	207
5.4.2 System complexity.....	211
5.4.2.1 Results.....	211
5.4.2.2 Discussion	215
5.5 Complexity and contact revisited	217
6 Concluding remarks and suggestions for further research	222
References and sources	233
Appendix	256

LIST OF FIGURES

Figure 1. Pronoun omission attestation across variety types.....	112
Figure 2. Pronoun omission attestation across regions	112
Figure 3. Residuals grouped by observed value	114
Figure 4. Pronoun omission pervasiveness across variety types	116
Figure 5. Pronoun omission pervasiveness across regions	117
Figure 6. Residuals and observed values.....	119
Figure 7. Object omission attestation across regions	121
Figure 8. Object omission pervasiveness across regions	122
Figure 9. Non-referential subject omission attestation across regions	123
Figure 10. Attestation of referential <i>it</i> omission across regions	124
Figure 11. Attestation of non-referential <i>it</i> omission across regions	126
Figure 12. Non-referential <i>it</i> omission pervasiveness across regions	126
Figure 13. Expression of pronominal subjects in WALS (F101A)	129
Figure 14. Proportion of omitted and overt pronouns per ‘Variety’	168
Figure 15. Proportion of omitted and overt pronouns per ‘Text type’	170
Figure 16. Proportion of omitted and overt pronouns per level of ‘Agreement’	171
Figure 17. Proportion of omitted and overt pronouns per level of ‘Accessibility’	173
Figure 18. Proportion of omitted and overt pronouns per level of ‘Coordination’	174

Figure 19. Proportion of omitted and overt pronouns per level of ‘Clause position’	176
Figure 20. Proportion of omitted and overt pronouns per level of ‘Sentence type’	177
Figure 21. Proportion of omitted and overt pronouns per level of ‘Priming’	179
Figure 22. Proportion of omitted and overt pronouns per level of ‘Clause type’	180
Figure 23. Proportion of omitted and overt pronouns per level of ‘Pronoun’	181
Figure 24. Median ‘Co-occurrence frequency’ of omitted and overt pronouns	183
Figure 25. Median ‘Collocational strength’ of omitted and overt pronouns	184
Figure 26. Proportion of omitted and overt pronouns per level of ‘Verb class’	185
Figure 27. Probability of omission per level of ‘Accessibility’	197
Figure 28. Probability of omission per level of ‘Priming’	199
Figure 29. Probability of omission per level of ‘Clause type’	199
Figure 30. Probability of omission per level of ‘Pronoun’	200
Figure 31. Probability of omission per level of ‘Verb class’	201
Figure 32. Probability of omission per level of ‘Text type’ in BrE, IndE, and SgE.....	203
Figure 33. Probability of omission as a function of the interaction between ‘Variety’ and ‘Coordination’	204
Figure 34. Probability of omission as a function of the interaction between ‘Variety’ and ‘Clause position’	205
Figure 35. ‘Verb form’ effects in BrE, IndE, and SgE	206
Figure 36. Variable importance in BrE, IndE, and SgE.....	214
Figure 37. Speaker preferences across varieties	297

LIST OF TABLES

Table 1. Analysis of example (28)	91
Table 2. Analysis of example (29)	92
Table 3. Attestation and pervasiveness in White Zimbabwean English	104
Table 4. Attestation and pervasiveness in White South African English.....	105
Table 5. Classification of varieties according to region and type	108
Table 6. Single term deletions from a model containing both variety type and region as predictors of pronoun omission attestation	113
Table 7. Single term deletions from a model containing both variety type and region as predictors of pronoun omission pervasiveness	118
Table 8. Classification of texts in <i>ICE</i>	137
Table 9. Texts included in the subsets of <i>ICE-GB</i> , <i>ICE-IND</i> , and <i>ICE-SIN</i>	139
Table 10. Contingency table for <i>I am</i> in IndE	157
Table 11. Variables included in the analysis	161
Table 12. Example of permutation in random forests	165
Table 13. Distribution of omitted and overt pronouns per ‘Variety’.....	168
Table 14. Distribution of omitted and overt pronouns per ‘Text type’	170
Table 15. Distribution of omitted and overt pronouns per level of ‘Agreement’	171
Table 16. Distribution of omitted and overt pronouns per level of ‘Accessibility’	172

Table 17. Distribution of omitted and overt pronouns per level of ‘Coordination’	174
Table 18. Distribution of omitted and overt pronouns per level of ‘Clause position’	175
Table 19. Distribution of omitted and overt pronouns per level of ‘Sentence type’	177
Table 20. Distribution of omitted and overt pronouns per level of ‘Priming’	178
Table 21. Distribution of omitted and overt pronouns per level of ‘Clause type’	180
Table 22. Distribution of omitted and overt pronouns per level of ‘Pronoun’	181
Table 23. Median ‘Co-occurrence frequency’ of omitted and overt pronouns.....	182
Table 24. Median ‘Collocational strength’ of omitted and overt pronouns.....	183
Table 25. Distribution of omitted and overt pronouns per level of ‘Verb class’	185
Table 26. Model summary of the binary mixed-effects logistic regression.....	190
Table 27. Results of the binary mixed-effects logistic regression	193
Table 28. Variance Inflation Factors.....	196
Table 29. Model summaries of BrE, IndE, and SgE random forests	212
Table 30. Variable importance in BrE, IndE, and SgE	213
Table 31. Model summary: variety type as predictor of pronoun omission attestation	256
Table 32. Contrasts: variety type as predictor of pronoun omission attestation	256
Table 33. ‘Model summary: region as predictor of pronoun omission pervasiveness	258
Table 34. Contrasts: region as predictor of pronoun omission pervasiveness	258
Table 35. Model summary: region as predictor of object omission attestation	261
Table 36. Contrasts: region as predictor of object omission attestation	261
Table 37. Model summary: region as predictor of object omission pervasiveness	264
Table 38. Contrasts: region as predictor of object omission pervasiveness	264
Table 39. Model summary: region as predictor of non-referential subject omission attestation	267

Table 40. Contrasts: region as predictor of non-referential subject omission attestation	268
Table 41. Model summary: region as predictor of referential <i>it</i> omission attestation	270
Table 42. Contrasts: region as predictor of referential <i>it</i> omission attestation	271
Table 43. Model summary: region as predictor of non-referential <i>it</i> omission attestation ...	274
Table 44. Contrasts: region as predictor of non-referential <i>it</i> omission attestation	274
Table 45. Model summary: region as predictor of non-referential <i>it</i> omission pervasiveness	277
Table 46. Contrasts: region as predictor of non-referential <i>it</i> omission pervasiveness	277
Table 47. Further ‘Variety’ contrasts	279
Table 48. Further ‘Text type’ contrasts	282
Table 49. Further ‘Coordination’ contrasts	288
Table 50. Further ‘Clause position’ contrasts	290
Table 51. Further ‘Accessibility’ contrasts	293
Table 52. Further ‘Priming’ contrasts	294
Table 53. Further ‘Verb class’ contrasts	295

ACKNOWLEDGEMENTS

This book is a revised version of my dissertation, submitted at the University of Santiago de Compostela in 2018. I am greatly indebted to my supervisors, Professors Teresa Fanego and J. Carlos Acuña-Fariña, for their helpful suggestions and encouragement during the planning and development of this project. I would also like to acknowledge the assistance and counsel received from Professors Bernd Kortmann and Benedikt Szmrecsanyi during my research stays at Albert-Ludwigs Universität Freiburg and KU Leuven. In addition, thanks are due to the research teams *Variation, Linguistic Change and Grammaticalization*, of which I am a member, and *Cognitive Processes and Behaviour*, and the research network *English Linguistics Circle* for their support. For generous financial support, I am grateful to the European Regional Development Fund and the following institutions: Regional Government of Galicia (Directorate General for Scientific and Technological Promotion, grants ED431B 2017/12 and ED431D 2017/09); Spanish Ministry of Innovation, Science and Universities (grants FFI2017-86884-P, FFI2014-52188-P and BES-2015-071233). Last but not least, this book would not have been possible without the constant help and encouragement of my partner, my family, and my friends.

LIST OF ABBREVIATIONS

Atlases and corpora

APiCS = Atlas of Pidgin and Creole Language Structures

eWAVE = Electronic World Atlas of Varieties of English

GloWbE = Corpus of Global Web-based English

ICE = International Corpus of English

ICE-IND = International Corpus of English (Indian component)

ICE-GB = International Corpus of English (British component)

ICE-SIN = International Corpus of English (Singaporean component)

WALS = World Atlas of Language Structures

Language status

L1 = First Language

L2 = Second Language

Periods in the history of English

OE = Old English

ME = Middle English

eModE = Early Modern English

PDE = Present-day English

Varieties of English

AborE = Aboriginal English

AmE = American English

AppE = Appalachian English

AusE = Australian English

AusVE = Australian Vernacular English

BahC = Bahamian Creole

BahE = Bahamian English

BelC = Belizean Creole

Bisl = Bislama

BlSAfE = Black South African English

BrC = British Creole

BrE = British English

ButlE = Butler English

CamE = Cameroon English

CamP = Cameroon Pidgin

CFE = Cape Flats English

ChcE = Chicano English

ChIsE = Channel Islands English

CollAmE = Colloquial American English

CollFijiE = Basilectal Fiji English

CollSgE = Colloquial Singapore English

EA = East Anglian English

EAAVE = Earlier African American Vernacular English

EMarC = Eastern Maroon Creole

FijiE = Acrolectal Fiji English

FlkE = Falkland Islands English

GhE = Ghanaian English

GhP = Ghanaian Pidgin

GuyC = Guyanese Creole

HawC = Hawai'i Creole

HKE = Hong Kong English

IndE = Indian English

InSAfE = Indian South African English

IrE = Irish English

JamC = Jamaican Creole

JamE = Jamaican English

KenE = Kenian English

LibSE = Liberian Settler English

MalE = Malaysian English

MltE = Maltese English

ManxE = Manx English

NfldE = Newfoundland English

NigE = Nigerian English

NigP = Nigerian Pidgin

Norfk = Norfolk Island/Pitcairn English

North = English Dialects in the North of England

NZE = New Zealand English

OSE = Orkney and Shetland English

OzE = Ozark English

PakE = Pakistani English

PalmE = Palmerston English

PhE = Philippine English

RAAVE = Rural African American Vernacular English

RRC = Roper River Creole

SanAC = San Andres Creole

Saram = Saramaccan

ScE = Scottish English

SE = English Dialects in the Southeast of England

SEAmE = Southeast Aeriana Enclave Dialects

SgE = Singapore English

SLkE = Sri Lankan English

StHE = Saint Helena English

SW = English Dialects in the Southwest of England

TdCE = Tristan da Cunha English

TorSC = Torres Strait Creole

TP = Tok Pisin

TrinC = Trinidadian Creole

TznE = Tanzanian English

UAAVE = Urban African American Vernacular English

UgE = Ugandan English

VinC = Vincentian Creole

VLibE = Vernacular Liberian English

WeE = Welsh English

WhSAfE = White South African English

WhZimE = White Zimbabwean English

1 INTRODUCTION

1.1 COMMUNICATIVE EFFICIENCY AND LANGUAGE COMPLEXITY

A recent paper (Selinger et al. 2015) published in *Current Biology* claimed that people are able to optimize their energy use in real time while they are walking so as to minimize the effort they have to make. Selinger et al. conducted an experiment in which they measured their subjects' step frequency, that is, how fast they moved, while they were walking on a treadmill. Subjects wore a light exoskeleton attached to their legs that penalized either high or low step frequencies by applying a force that restricted the movement of their knee joints: a penalize high function applied force when their step frequency was high (i.e., when they walked fast) and a penalize low function when it was low (i.e., when they moved slowly). Selinger et al. found that subjects modified their preferred step frequencies in the face of both high and low step frequency penalizations to achieve the minimum level of energy use possible. They did this consistently and rapidly – only a few seconds after the penalization was applied – and even when very minor reductions in energetic cost were at stake, in the range of 4% to 8%.

What these findings reflect is an efficient exploitation of energy resources on the part of the subjects that participated in Selinger et al.'s (2015) experiment. Sometimes, achieving the optimum level of energy expenditure required by the task at hand meant that subjects had to walk more slowly (i.e., when the exoskeleton penalized high step frequencies). On other

occasions, however, this meant walking faster (i.e., low step frequency penalizations), even if subjects initially had to make a greater effort to avoid the penalization of the exoskeleton.

Similarly, a tradition of linguistic research going back to Zipf (1949) argues that human communicative behaviour is also guided by an efficient use of energy. Zipf's principle of least effort, which postulates that language users prefer to achieve their immediate and future communicative goals with the least amount of effort required to do so, entails a trade-off between the opposing preferences of speakers, who benefit from less articulatory effort, and hearers, who favour transparency. In the domain of the lexicon, Zipf found a balance between word frequencies and number of words employed since, according to his principle, speakers would prefer fewer high-frequency words, while hearers would favour many low-frequency words. This is captured in the fact that a word's frequency multiplied by its frequency rank is almost always constant, which in turn entails that the most frequent word occurs approximately twice as often as the second most frequent word, three times as often as the third most frequent word, and so on. In the domain of morphosyntax, Hawkins (1994, 2004, 2009, 2014) shows that performance and grammar are highly correlated, in the sense that structures that facilitate language processing relative to the minimum effort required for an efficient transmission of information tend to become part of the grammars of languages. This idea is articulated in his "Performance-Grammar Correspondence Hypothesis" (Hawkins 2004, 3):

Grammars have conventionalized syntactic structures in proportion to their degree of preference in performance, as evidenced by patterns of selection in corpora and by ease of processing in psycholinguistic experiments.

Hawkins' concept of communicative efficiency also implies a balance between the preferences of speakers and hearers: sometimes, in order to achieve a successful transmission of information, the speaker has to make a greater effort but, on other occasions, this might not be necessary. For instance, when referring to an entity that can be easily retrieved from memory by the addressee, it might suffice to use a short and ambiguous form, such as a personal pronoun. If the antecedent is not accessible enough, however, longer referential expressions, such as a complex noun phrase with an embedded relative clause, might have to be resorted to.

More predictable, and, therefore, less informative, words are usually pronounced with a shorter duration and less articulatory detail than less predictable/more informative words (Jaeger and Tily 2011). Tily and Piantadosi (2009) show that predictable antecedents are referred to by means of shorter and more ambiguous expressions, such as pronouns, and that the probability of using a longer and more specific referential expression increases as the antecedent becomes less predictable. These patterns reflect an efficient use of energy resources on the part of language users whereby more articulatory effort, and thus energy, is spent when this is required for the successful transmission of information, but costs are also minimized whenever possible.

A necessary characteristic of any communicative system that strives to be efficient is the possibility of leaving certain pieces of information underspecified when the context is informative enough to disambiguate (Piantadosi, Tily, and Gibson 2012). That is, an efficient communicative system has to allow the use of vague, ambiguous, and/or zero-specified forms under certain conditions to convey information that is predictable, easily retrieved from the context, or unimportant for the successful transmission of the message (cf. Hawkins 2004, 40). By doing this, speakers do not have to waste energy articulating forms that express

unnecessary information that, for instance, can already be inferred from the context, which in turn entails a reduction in redundancy. There is an extensive body of research that suggests that hearers are highly skilled in using contextual information to disambiguate and fill in the gaps left in the message by speakers (cf., for instance, Piantadosi, Tily, and Gibson 2012, and references therein). Efficient communicative systems, therefore, should exploit the inferential abilities of the hearer whenever possible as “inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference” (Levinson 2000, 29). Naturally, miscommunication sets the limits for the underspecification of information, as this is terribly inefficient for both types of language users (cf. Wasow, Perfors, and Beaver 2005): any potential cost reduction achieved by speakers by means of using vague, ambiguous, or zero-specified forms is outweighed by having to explain the intended meaning of the structure to the hearer or having to articulate the message again.

A notion related to that of communicative efficiency is language complexity. The concept of efficiency entails that there are some forms or structures that are simpler and easier to process than others and that these forms will be used in those contexts in which a reduction in complexity does not endanger the successful transmission of the message. Language complexity is a topic that has been dealt with by numerous scholars going back to (at least) Hockett (1958) and that has played a role, even if only as a background assumption, in many different research traditions. Initially, most linguists assumed that all languages were equally complex but, nowadays, it has been demonstrated that they may differ in this respect and that several factors foster variation in language complexity. A factor that has figured prominently in the specialized literature, and one that is also considered in this monograph, is language contact. Contact with other languages or dialects has been shown to result in grammatical

simplification, although the situation is not as simple as that, as the results of the case studies presented in Chapters 4 and 5 will make evident.

1.2 AIMS AND RESEARCH QUESTIONS

Against this theoretical background, the main goals of the present book are to assess the complexity/simplicity and efficiency/inefficiency of pronoun omission and to examine how contact affects the distribution of pronoun omission features in varieties of English around the world. Pronoun omission, that is, the presence of a gap in a sentence that could have been filled by, usually, a personal pronoun, generates simpler structures with fewer forms and, consequently, less articulatory effort on the part of the speaker. This reduction in formal complexity is not necessarily at the expense of the hearer, who should still be able to successfully decode the message without having to incur in extra processing costs due to the fact that the antecedents of omitted pronouns are (almost) exclusively highly predictable and accessible referents that can be easily retrieved from contextual information. Simpler structures with fewer forms for the speaker to articulate that can still be decoded by the hearer should, in principle, allow for a more efficient transmission of information, assuming that inference is indeed cheap.

The present work focuses on pronoun omission features in different dialects or varieties of English so as to assess their contribution to the complexity and communicative efficiency of their grammars. This linguistic phenomenon is particularly well-suited to this purpose for several reasons. First, it is a reasonably frequent feature if one adopts a sufficiently broad definition of pronoun omission. This enables the researcher to retrieve a sufficiently large data sample from, for instance, corpus materials so as to be able to reach generalizable conclusions. Second, producing a pronoun in omitted or overt form involves no change in

meaning if one does not consider the discourse status of antecedents.¹ To illustrate this, consider example (1):²

(1) a. $\emptyset$ Hope you are well.

b. I hope you are well.

No noticeable change in meaning can be appreciated between (1a) and (1b), which means that we are dealing with an (almost) purely formal alternation between two largely semantically equivalent variants. This allows us to focus on the mechanisms underlying the formal reduction found in (1a) in comparison to (1b) without the potential interference of semantic variables. Finally, pronoun omission is a cross-linguistically well-researched grammatical feature, even in a language such as English, where it is a highly restricted phenomenon. Therefore, the factors that influence the choice between omitted and overt pronouns are fairly well understood.

The World Englishes paradigm (cf., for instance, Mesthrie and Bhatt 2008) was considered to be a particularly relevant and appropriate framework for the purposes of the present monograph in order to evaluate the general claim that language contact simplifies grammars. Many scholars have applied the notion of complexity to the analysis of variation in English with the aim of “understanding the comparatively simple (i.e. language-internal complexity variation) before approaching the comparatively complicated (i.e. cross-linguistic

¹ One could argue that speakers use different referential expressions to signal to the addressee the level of mental accessibility of the referent. Under this interpretation, there would be a change in meaning from (1a) to (1b), namely that the antecedent of the omitted pronoun in (1a) is estimated by the speaker to be more accessible for the addressee than that of (1b).

² $\emptyset$ is used throughout this book to mark the position of the omitted pronoun in the clause if it had occurred in an overt form.

complexity variation)” (Szmrecsanyi and Kortmann 2012, 14). The study of varieties of English is a promising field of research to understand the connection between contact and complexity or efficiency, since the spread of English around the world has caused the emergence of new varieties in the former British (and American) colonies, often as a result of different types of language contact. Given that this has been a relatively recent phenomenon, information regarding the sociohistorical conditions in which these new varieties developed is readily available, and this in turn means that we can establish connections between social determinants and their putative effects on grammatical complexity. These connections can then be extrapolated to and tested in other languages for which sociolinguistic data are not so accessible. Assuming that contact results in simplification, pronoun omission should be more frequently attested in varieties exposed to some degree of language contact, and, as shown in Chapter 4, this is indeed the case.

Some other general research questions that are addressed in the present book include the following:

1. When do speakers of (varieties of) English omit pronouns? Is this done at random or in an efficient manner, that is, when it does not hinder the successful communication of information from the speaker to the hearer?
2. What are the language-external and language-internal constraints that influence the choice between omitted and overt pronouns in English? What consequences do they have for the complexity and efficiency of the grammars of the varieties under study?

Research question one addresses the issue that, as argued above, pronoun omission only entails an increase in communicative efficiency in those cases in which it does not cause

processing difficulties for the hearer. Omitted pronouns must refer to highly accessible antecedents, otherwise the hearer may not be able to coindex the omitted pronoun with its correct referent, in which case communication fails and the speaker may be compelled to articulate the message again or, at least, clarify which was the entity being alluded to. In addition, more formally explicit variants may be preferred in certain structural contexts for reasons of processing ease. Research question two, on the other hand, enquires into the pronoun-deletion grammar(s) of (varieties of) English. Language complexity and communicative efficiency are also affected by the number of constraints that regulate the omission of pronouns, since a larger number of rules would result in a more complex and, in principle, less efficient grammar.

1.3 STRUCTURE OF THE BOOK

This monograph is structured as follows. Chapters 2 and 3 provide the theoretical background necessary to understand and evaluate the findings of the case studies presented in Chapters 4 and 5. Chapter 2 reviews the specialized literature on language complexity, starting with a historical overview on the evolution of the concept and its importance in linguistic theorizing. Then, we examine previously identified sources of complexity variance, paying special attention to the simplifying effect of contact on grammatical complexity and possible reasons for this influence. Certain metrics of complexity are subsequently inspected as examples of the manner in which previous scholars have approached the issue of how to measure the complexity of languages. Some important complexity notions and distinctions are extrapolated from these metrics and the need to investigate them separately is emphasized. First, we need to distinguish between absolute and relative approaches to the study of complexity as they provide different, and sometimes conflicting, assessments: whereas

absolute complexity metrics consider the complexity of languages in a vacuum as abstract systems, relative metrics take into account the preferences of language users in their estimations. Second, it is essential to specify the nature of the objects that are measured, since there are fundamental theoretical and empirical differences between considering the complexity of particular linguistic structures or that of grammatical rules or paradigms. Finally, several applications of the notion of complexity to the field of World Englishes are reviewed.

Chapter 3 considers the literature on pronoun omission from a cross-linguistic perspective so as to isolate the factors that constrain the occurrence of omitted and overt pronouns across languages. Generative approaches have dealt extensively with this linguistic phenomenon, highlighting the role of agreement between the omitted argument and the verb as a licenser of pronoun omission: omission occurs when agreement morphology is strong enough to identify the morphological features of the omitted argument. However, as pronoun deletion features are also attested in languages with poor or no agreement morphology, other syntactic explanations have also been postulated to account for the occurrence of pronoun omission in, for instance, isolating languages such as Mandarin Chinese. More cognitive approaches, on the other hand, argue that pronoun deletion features cannot be accounted for only in purely syntactic terms. Despite the importance of agreement morphology, the role of the context, among other factors, must also be taken into consideration in order to explain the occurrence of omitted pronouns. Then, the chapter zooms in on pronoun omission in English. A brief overview on the evolution of pronoun omission in the history of British English is presented, followed by an explanation of the constraints regulating the choice between omitted and overt pronouns in Present-day English that have been put forward in previous research. Finally,

Chapter 3 includes an analysis of pronoun deletion in terms of its contribution to the relative structural and system complexity (and efficiency) of English.

Chapter 4 contains the first of the two case studies in the monograph. In this chapter, the results of a cross-varietal study are presented which quantifies the attestation and pervasiveness of pronoun omission features in varieties of English using data from *The Electronic World Atlas of Varieties of English* (Kortmann and Lunkenheimer 2013). The main goal of this study is to provide a bird's-eye perspective on the distribution of pronoun omission features across varieties and to assess the influence of language contact. The explanatory power of two factors is examined: variety type, functioning as a proxy for the simplifying effect of second-language acquisition and use, and region, a proxy for substrate language influence. The results of this study support the claim that pronoun omission generates simpler and more efficient structures, as high-contact varieties of English exhibit a higher attestation rate of pronoun deletion features than low-contact varieties. In addition, Asia and the Pacific emerge as the omission regions par excellence, thus suggesting that substrate languages may indeed (dis)favour the occurrence of omitted pronoun in particular varieties.

The second case study is the focus of the next chapter. To complement the results of the cross-varietal survey included in Chapter 4, a corpus-based study of referential subject pronoun omission in British English, Indian English, and Singapore English is conducted in Chapter 5 with data from the *International Corpus of English*. This type of analysis sheds light on fine-grained usage patterns of omitted and overt pronouns in those three varieties, which differ in terms of their degree of contact with other languages and dialects, their statuses as first- or second-language varieties, and the linguistic ecologies and sociohistorical conditions in which they developed. The influence of several language-external and language-

internal constraints on the choice between omitted and overt pronouns is explored with the aim of empirically testing the conclusions on the relative structural and system complexity of pronoun omission reached in Chapter 3.

Finally, Chapter 6 concludes with some generalizations that can be extracted from the studies included in Chapters 4 and 5 of this book. In addition, some potential avenues for further research are considered.

2 LANGUAGE COMPLEXITY

Language complexity has occupied a central position in linguistic discussion from the very first half of the twentieth century, and it was invoked as an explanatory factor even as early as the nineteenth century. However, despite its importance in linguistic theorizing, it was not until the end of the twentieth and beginning of the twenty-first centuries that complexity was taken seriously and approached in a systematic manner. Metrics of complexity have been proposed and the old assumption that “all languages are equally complex” (Deutscher 2009, 243) has been rejected on the basis of actual empirical data. As a first approximation to the topic, we can quote the definition provided by Rescher (1998, 1; quoted in Karlsson, Miestamo, and Sinnemäki 2008, VIII): “Complexity is first and foremost a matter of the number and variety of an item’s constituent elements and of the elaborateness of their interrelational structure, be it organizational or operational.” This definition captures the intuition that a given object, a language in this case, is more complex if it comprises a large number of elements and different types of elements, and if a large number of rules are necessary to describe how these elements are structured and put into use. These ideas figure, in one way or another, in most metrics of complexity put forward to date. Differences between metrics can be found, however, in which elements (and types of elements) are measured in each one, and in the perspective from which they approach these elements (cf. Section 2.3 below).

The remainder of this chapter is structured as follows. Section 2.1 provides the historical background of the concept of language complexity, focusing on the assumption that languages do not vary with respect to the complexity of their grammars, prevalent in twentieth-century linguistics, and on recent challenges to this idea. Section 2.2 revolves around language contact and its influence on grammatical complexity. Section 2.3 describes some important concepts which must be considered when measuring the complexity of grammars. Finally, in Section 2.4, a range of studies dealing with complexity in varieties of English, one of the research fields to which this monograph is circumscribed, is presented.

2.1 HISTORICAL BACKGROUND

In the nineteenth century, linguists and philosophers considered language to be the means of expression of a people's or a nation's spirit. Therefore, differences between languages were ultimately assumed to have their source in differences between their speakers as regards their character or even their mental capacities (Von Humboldt 1836; quoted in Szmrecsanyi and Kortmann 2012, 7). On the basis of this framework, if a language was simpler than another, this meant that speakers of the former had in some way a simpler set of capacities than those of the latter. Indo-European languages, with their complicated morphological systems, were thought to be the most complex, and thus perfect, languages, able to convey the abstract kind of ideas that is characteristic of developed societies. This theory, however, was eventually abandoned because it did not fit well with more egalitarian ideologies that appeared in the course of the twentieth century concerning human nature: all humans were now thought to be biologically identical, with comparable physical and mental abilities, and, therefore, with the same fundamental rights. Against this backdrop, the assumptions that lay at the core of linguistics also changed, and racist ideas about language complexity were rejected.

One of the first important linguistic movements of the twentieth century was descriptivism, founded by Franz Boas and Leonard Bloomfield (cf. Sampson 2009). Rooted in the egalitarian spirit of the century, they cast aside nineteenth-century ideas and set out to prove that the languages of ‘primitive’ communities were at least as complex as those spoken in Western societies. This led to the belief that, since all humans are inherently equipped with the same set of mental, cultural, and physical capacities, they must all speak languages that were comparable in terms of their overall complexity. One of the earliest explicit formulations of this belief is the by now famous quote by Hockett (1958, 180–81):

Objective measurement is difficult, but impressionistically it 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. Fox, with a more complex morphology than English, thus ought to have a somewhat simpler syntax; and this is the case.

Simply put, the assumption was that complexity was a fixed feature of human languages and, therefore, all languages were equally complex because a decrease in complexity in one grammatical domain was compensated by an increase in another domain. The justification for this belief was that all languages had the same communicative function, and so the lack of some grammatical mechanisms in, say, the morphological component had to be balanced out by further complexities in, for instance, syntax, an idea that is nowadays known as “the trade-off hypothesis” (Sinnemäki 2008, 68). It is worth noting that, as can be deduced from the very first line of Hockett’s quote, “Objective measurement is difficult, but impressionistically it

would seem [...],” no empirical evidence was ever provided to support the assumption of complexity invariance. Rather, this belief, which went on to become a central tenet of twentieth-century linguistics, was based on ideological motives (Sampson 2009, 4): one of the goals of the descriptivist school was to demonstrate that all languages had intricacies in their grammars that made them complex, and thus that all humans had the same mental and cultural abilities regardless of how technologically developed the community was in which they lived.

After descriptivism came generativism, and with it the very conception of language changed. Descriptivists believed that language was a cultural artefact that could vary in every way imaginable (except in overall complexity) and evolve to be adapted to the needs of its speakers. For generativists, however, language was not part of human culture but of human biology: we are born with an innate universal grammar that underlies all human languages and, therefore, they are all basically identical in terms of structure, except for minor parametric variations (Chomsky 1980, 134, 1991, 26). One premise did remain constant between descriptivism and generativism, however, the assumption that all languages were equally complex. As mentioned above, for descriptivists this belief was grounded in ideological motivations: it was a way to prove that all humans were equal, since they all spoke complex languages. In generativism, this assumption derived from the very theoretical foundations of the school, since the assumed innate cognitive machinery dedicated to language was too comprehensive to allow for differences in terms of grammatical complexity, neither from a synchronic nor from a diachronic perspective (Sampson 2009, 7). Additionally, individuals were also considered to have essentially the same competence in their mother tongue, and this was assumed to be true both between individuals and within the same individual throughout her or his life.

At the turn of the twenty-first century, many scholars challenged the assumption of complexity invariance on theoretical, methodological, and empirical grounds (cf. Miestamo, Sinnemäki, and Karlsson 2008). The following is a short selection of studies that provide evidence against this assumption, and that have effectively demonstrated that we should consider as the null hypothesis that grammatical complexity varies between and within languages.³

From a theoretical standpoint, Deutscher (2009) criticizes two premises that underlie the assumption, which he terms the ‘minimum argument’ and ‘the maximum argument.’ The minimum argument suggests that languages must have equal levels of grammatical complexity due to the fact that they are all used for the same essential purpose of communicating information (cf. Hockett’s quote above). This implies that they all must have some resources to fulfil this function, which imposes a minimum level of complexity on their grammars. However, languages can be more complex than this minimum level required for communicating information successfully, which indicates that, in theory, they can differ with respect to their overall complexity. The maximum argument suggests that there is a limit in the grammatical complexity that any language can have, due to issues such as cognitive restrictions on learnability and/or language processing. Most languages (except pidgins and creoles) have had time to reach that maximum level of complexity, and thus, it is argued, they must be equally complex. However, as Kusters (2008) contends, the human brain is able to learn and process more than one linguistic system, which means that no individual language can exceed the maximum level of complexity that the brain can hold.

³ For further challenges to the assumption of invariant complexity see, for instance, the collection of papers in Miestamo, Sinnemäki, and Karlsson (2008) or Sampson, Gil, and Trudgill (2009).

Miestamo (2006b, 2008) argues on methodological grounds that metrics of overall complexity face two problems related to their representativity and comparability. First, whereas such metrics must account, in principle, for all aspects of a grammar in order to measure overall complexity, in practice it is very difficult, or even impossible, for linguists to take into account all relevant considerations, which decreases representativity. Second, any global metric of complexity must measure all the subsystems of a grammar, but these subsystems contain elements of a different nature. This means that, methodologically speaking, it is not possible to compare the results obtained for each of them. Taken together, these two problems make any discussion of overall grammatical complexity virtually meaningless unless a metric is devised that can overcome such limitations.

The assumption of invariant complexity has also been challenged on empirical grounds. Fenk-Oczlon and Fenk (2008) compare the complexity of eight Indo-European languages and conclude that balancing effects between the different subsystems of their grammars exist, so that, for instance, high phonological complexity correlates cross-linguistically with low morphological complexity, and vice versa. However, these balancing effects do not support the trade-off hypothesis, because, even if languages with a more complex phonology do tend to have a simpler morphology, this does not mean that, for each language, the sum of the complexities of the phonological and morphological subsystems is always the same. Fenk-Oczlon and Fenk (2008) suggest that it was these cross-linguistic balancing effects that gave rise to the trade-off hypothesis in the first place, but they are not enough to accept it on empirical grounds. Sinnemäki (2008) approaches the hypothesis from a different perspective. He studies the functional domain of core argument marking and looks for trade-offs between different grammatical strategies: if arguments are marked morphologically in a language, they should not be marked by word order or other mechanisms (provided, of course, that the

hypothesis holds).⁴ Using a sample of fifty languages, he finds that, cross-linguistically, there is a small negative correlation between morphological marking and word order but that there are also languages in which both strategies are used to distinguish between different arguments of the verb. This means that, even though there may be balancing effects between different marking strategies within the same functional domain, there are no trade-offs, because some languages have both a more complex morphology and a more complex syntax, that is, rigid word order, than others.

The conclusion that we can extract from these studies is, then, that the assumption of invariant complexity does not hold on empirical grounds, since systematic analysis of data points to the fact that less complexity on a particular grammatical domain is not necessarily compensated by more complexity on a different one. Additionally, there exist no compulsory minimum nor maximum level of complexity that languages must reach, so no a priori theoretical reason to assume that all languages must be equally complex. Finally, the task of measuring the overall grammatical complexity of a language is not viable on methodological terms, because no metric can consider every factor that should be accounted for nor compare complexity ratings between different grammatical domains. Nowadays, most scholars agree that languages do not have to be equal with respect to their overall grammatical complexity, and research has gravitated towards finding the factors underlying complexity differences between and within languages.

2.2 SOURCES OF COMPLEXITY VARIANCE

⁴ A functional domain is defined as “a set of closely related semantic or pragmatic functions that are linguistically encoded by at least some languages” (Sinnemäki 2008, 68).

While it is true that there is some level of complexity at which many languages seem to converge,⁵ there exist significant differences between them in terms of their overall degree of grammatical complexity, and even more if we focus on particular domains of grammar (Nichols 2009). The question now is why languages differ in this respect. One of the best-known sources of complexification in grammar is time: with the passage of time, languages generate new grammatical distinctions and categories through grammaticalization and other “maturation” processes which presuppose a number of developmental stages and, therefore, time to go through all of them (Dahl 2004, 2). These processes all entail an increase in the grammatical mechanisms and devices available in the grammar of a language, which means that they result in an increase in complexity.⁶ In this connection, Deutscher (2000) shows that the grammar of early Akkadian, one of the first languages to have been written, did not have the mechanisms to form finite complement clauses, but that, in its 2,000-year history, this kind of structures gradually developed out of simpler main-clause devices that existed before. Deutscher (2000) claims that this growth in syntactic complexity in Akkadian was connected to new communicative needs that arose in its speech community.

The passage of time is also a source of complexity at a smaller scale. Sampson (2001), in a corpus study of clause subordination in spontaneously spoken British English, found that older speakers used on average a larger number of subordinated structures than younger ones, so that, for instance, thirty-year-old speakers used fewer of these structures than forty-year-

⁵ Nonetheless, this seems to be mostly because, as with many other phenomena, languages tend to follow a normal distribution with respect to their complexity (Nichols 2009), that is, the majority exhibit an average level of complexity, with only a few languages being more or less complex than this average.

⁶ Other diachronic changes, such as those that are the result of, for instance, analogical processes, can be considered to simplify grammar. However, in the sociolinguistics literature on complexity, it is suggested that, when this type of changes dominate in the history of a language, this is not a ‘natural’ linguistic process but rather a consequence of language contact (cf. Trudgill 2009a; and Section 2.2.3).

old ones, and, in turn, these used fewer subordinate clauses than sixty-year-olds. This seems to suggest that complexity, understood in this case as the incidence of recursive devices in language use, also grows in an individual's lifetime.

Other aspects also cause differences in complexity at the individual level. Dąbrowska (1997) and Chipere (2003, 2009) demonstrate that factors such as the educational level of the speaker and whether she or he received formal instruction in a language influence the ability to process syntactically complex structures with several degrees of embedding. For instance, individuals with a lower educational level had more problems to comprehend the sentence *The doctor knows that the fact that taking care of himself is necessary surprises Tom* than those with a higher one (Dąbrowska 1997). Additionally, speakers who received explicit instruction in the language also performed better than those who did not (Chipere 2003, 3, 2009).

2.2.1 Language contact

Leaving individual differences aside and going back to the language level, a source of complexity variance that figures prominently in the specialized literature, and a very important one for the purposes of the present work, is language contact. Language contact, and the concomitant influence of the processes of second-language acquisition and use, is a well-known factor causing grammatical simplification (cf., among others, Karlsson, Miestamo, and Sinnemäki 2008; Newmeyer and Preston 2014; Sampson 2009; Szmrecsanyi and Kortmann 2012). The main idea is that, a language, at some point in its history, may become used as a second or vehicular language by the majority of its speakers within a particular speech community. Consequently, in the process of learning and using this language, speakers will dispense with a vast amount of its grammatical machinery that is in

some way difficult for them to acquire and/or use. This will result in a simplified system – with, among other characteristics, fewer semantic/pragmatic categories being overtly coded, fewer irregularities, and fewer rules mapping meanings to forms – in response to the needs of its second-language users. Klein and Perdue (1997) clearly illustrate the point that second-language learning and use induces the loss of grammatical complexity in a language. In a cross-linguistic study following 40 learners of different languages for a period of time, Klein and Perdue found that all of them, at some point, reached a phase in which they used a relatively stable linguistic system, labelled the Basic Variety. This system, which is governed by a small set of principles and is largely unaffected by the specific interactions between source and target language, is mostly characterized by being simple but, at the same time, efficient, that is, enough to cover the learners’ basic communicative needs. Klein and Perdue (1997, 303) see the Basic Variety as representing “a particularly natural and transparent interplay between function and form in human language” that reflects the more essential aspects of human communication, and fully developed languages as being complexifications of this basic form.⁷ Most of the learners in Klein and Perdue’s study went on to develop their Basic Variety into more complex systems. However, about one-third of the subjects became fossilized in this stage while still being able to transmit information in an efficient manner in their second language.

John McWhorter is one of the fiercest advocates of the simplifying effects of contact. He basically initiated the complexity debate in the twenty-first century with his 2001 paper on the complexity of creoles, in which he claims that these languages (and, therefore, pidgins) have the simplest grammars of all. McWhorter (2001) argues that older languages are riddled with

⁷ This is not to be understood as meaning that fully-fledged languages developed from some basic blueprint, but rather that their grammars instantiate a more complex and elaborated version of that of the Basic Variety while, at the same time, being comparable with it in terms of the set of organizational principles that underlie both systems.

complexities, which arose with the passage of time by means of grammaticalization and other related processes. These complexities give overt expression to grammatical distinctions which serve a function in their particular grammars but are unnecessary for efficient communication, a fact evidenced by their absence in many of the world's languages. Most languages, however, have developed this kind of complexities by virtue of having been passed down uninterruptedly from generation to generation for several millennia, and as a result of the natural tendency of systems to become more complex over time (McWhorter 2001, 131). Creoles, on the other hand, originated from pidgins, that is, basic systems with only the essential mechanisms for effective communication, and have merely existed for five or six centuries. Therefore, they have not had enough time to generate the type of grammatical complexities characteristic of older languages, which is evident, for instance, in the absence in creoles' grammars of elaborated inflectional paradigms. Moreover, if we compare them with older isolating languages, the latter are still much more overspecified despite lacking the morphological intricacies of their inflected and agglutinative counterparts.

According to McWhorter (2007), however, the simplifying effects of contact are not only manifested in 'young' languages such as pidgins and creoles. Some older languages also display simplified grammars that could only have been the result of imperfect acquisition and use by vast numbers of adult second-language speakers. These include languages such as English, Mandarin Chinese, Persian, and Malay, among others, which at a certain point in their histories have entered into contact with other languages and have thus been subjected to extensive non-native acquisition. English, for instance, is the only Indo-European language without grammatical gender, and the only Germanic one without verb-second word order, a second-person singular pronominal form, or directional adverbs such as the earlier forms *hither* and *thither* (McWhorter 2008, 168).

In McWhorter's theory of grammatical simplification, language contact plays a very important role: it is not only one source of complexity reduction among many in the history of languages, but the only possible one. He postulates the following hypothesis for the relation between contact and simplification:

In the uninterrupted transmission of a human language, radical loss of complexity throughout the grammar is neither normal, occasional, nor rare, but **impossible**. The natural state of human language is one saddled with accreted complexity unnecessary to communication. Wherever this complexity is radically abbreviated overall rather than in scattered, local fashion, this is not just sometimes, but **always** caused by a sociohistorical situation in which non-native acquisition of the language was widespread enough that grammar was transmitted to new generations in a significantly simplified form (McWhorter 2008, 169; emphasis in original).

According to this hypothesis, with the passage of time languages undergo simplification processes in some grammatical domains, but these are not enough to produce a radically simplified system because they also tend to develop new complexities over time if transmitted in their complete form from one generation to the next. If a grammar displays a dramatically simple form, this can only be the result of a past (or present) in which second-language users constituted a high percentage of its speakers within a particular speech community. This is a very strong claim that requires substantiation (cf. McWhorter 2008 for some of examples), and any language that appears to be radically simple without having been exposed to a high degree of non-native learning and use poses a problem for this theory (cf. Gil 2008, 2009 for a potential counterexample). Therefore, for the purposes of the present book, I will not commit

to such a strong claim but treat contact as one source of complexity variance among other possible ones.

2.2.2 Acquisition versus use

Grammatical simplification due to contact has traditionally been hypothesized to be caused by non-native adult language *acquisition*, that is, in the process of learning their second language, adult and adolescent speakers discard (consciously or unconsciously) those grammatical features and mechanisms that are too difficult or complex for them to acquire (Trudgill 2009a). This means that features that include irregularities, new semantic/pragmatic distinctions into the grammar, or distinctions that are not transparently coded are usually cast aside by learners (cf. Kusters 2003, 2008; Miestamo 2008). However, there remains the issue of the effects of second-language *use* rather than acquisition in situations of language contact. Filipović and Hawkins (2013) highlight the need to consider multiple factors when trying to understand the process of second-language learning, including transfer from the first to the second language, general constraints on learning (that is, acquisition), processing limitations in production and comprehension (that is, use), and social and pedagogical factors such as motivation and teaching material. On the basis of previous models that take a series of principles into consideration (cf., among others, Ellis 1998; Larsen-Freeman 1997; MacWhinney 2005; Slobin 1977), they propose a comprehensive model for second-language acquisition that focuses on individual factors (transfer, and learning and processing constraints) rather than on more social or external circumstances. Their point of departure is the assumption that learners, as in the case of first language users, strive to communicate efficiently in their second language, that is, they try to transmit information to the addressee as fast as possible and with the minimum effort required by the communicative situation

(Hawkins 2009, 253; cf. Section 2.3.3). Filipović and Hawkins argue that any theory of language learning must take into account considerations related to first language influence on the second language and, more importantly for our purposes, both acquisition and production/comprehension processes. They propose a model that includes four general principles that guide the process of second-language learning. These guidelines, stated as a series of second-language user preferences, are the following (Filipović and Hawkins 2013, 158–59):

- **Minimize Learning Effort (MiL):** second-language users try to reduce the effort required to learn the target language, for instance, by transferring those features that are shared by their first and second languages, and by focusing on those features that occur frequently and that are structurally simple.
- **Minimize Processing Effort (MiP):** second-language learners try to reduce the processing effort required to use the grammatical features of the target language. The same situations that reduce learning effort apply in this case, that is, processing effort is minimized by means of transfer from the first to the second language, because the same mechanisms to process the first language can be used for the second one, and when grammatical features are simple and occur frequently in the input.
- **Maximize Expressive Power (MaE):** second-language learners try to express the same ideas and messages that they would in their first language. This is done by increasingly using more complex structures, and by allowing the transfer of features from the first language, both when they are shared with the second language and, in some cases, when they are not.

- **Maximize Communicative Efficiency (MaC):** second-language users try to increase communicative efficiency when transmitting information, that is, by reducing the time employed in conveying the message to the addressee and minimizing the processing effort necessary to do so. This means that the learner will need to use more complex structures in some situations and, in others, she or he will be able to resort to simpler ones.

These four principles interact with one another, sometimes with a reinforcing effect and, on other occasions, with a contradicting one, to generate the output of second-language users. For instance, whereas MiL and MiP are basically principles of least effort that describe the learners' preference for more economical paradigms and structures, or for transferred features that ease acquisition and use, MaE and MaC sometimes have the opposite effect of increasing complexity in those cases in which it is necessary for expressive purposes. It is important to note that learning and processing are kept separate, because, even though their effects tend to converge, they sometimes work in different directions. For instance, acquiring irregular constructions, such as the past tense and participle forms of irregular verbs in English, is difficult and requires an effort for second-language users, but, once they are learned, they are usually processed easily due to their high frequency. On the other hand, a transparent one-to-one mapping between meanings and forms makes acquisition less costly but results in an increased processing effort, since shorter forms which depend on the context for interpretation are typically easier to produce (Filipović and Hawkins 2013, 153; cf. also Hawkins 2004, 2009, 2014). Filipović and Hawkins' model postulates that second-language learners tend to minimize the effort required both to acquire and to use their second language.

Williams (1987, 1988, 1989) also claims that learner varieties, within which she includes foreign-language varieties and “non-native institutionalized varieties of English” (Williams 1987, 162), such as the ones spoken in India and Nigeria, are influenced by both acquisitional and usage considerations. On the one hand, learners tend to avoid inherent vulnerabilities in the grammar of a language, which means features that pose problems for them. On the other hand, learners are also guided by contradictory production principles: a tendency to increase economy, including the production of regular and minimally redundant structures (Williams 1987, 169), and a preference for reducing ambiguity by maximizing transparency and saliency (Williams 1987, 178). Crucially, Williams argues that the first principle, economy, reflects the inclination of speakers to keep production costs to a minimum, that is, to articulate structures that are as simple and economical as possible. On the contrary, increasing transparency and salience works in favour of the listener, since reducing ambiguity makes structures easier to decode and comprehend.⁸ These two principles influence the production of second-language learners, who, due to their restricted competence in the target language and processing limitations in real-time conversation, tend to produce maximally economical structures without redundant elements, which still comply with the listener’s need for clarity (Williams 1989, 365). Economy and the reduction of ambiguity are also essential to understand the presence or absence of certain grammatical features in contact varieties of English (Williams 1987; cf. Section 2.4).

What the studies reviewed in this section emphasize is the need to keep acquisition and use separate because, even though they do tend to converge, they may exert opposing pressures. This distinction is crucial for the purposes of the present work since, as will be

⁸ Increased explicitness may also help speakers monitor their own output, especially in cognitively complex environments (Williams 1987, 179).

argued in Section 2.3.5 and in Chapter 3, Section 3.4, pronoun omission can be seen as either increasing or reducing complexity depending on the perspective from which we approach this phenomenon, that is, from the point of view of acquisition or that of use, among others.

2.2.3 A typology of contact situations

Most studies on complexity within the field of sociolinguistics agree that language or dialect contact leads to grammatical simplification. On the contrary, in the typological tradition, the most widely accepted view is the opposite one, that is, that contact triggers an increase in grammatical complexity (cf. Heine and Kuteva 2005, 171; Trudgill 2011, 26–27). How can these two contradictory sets of findings be reconciled? Trudgill (2009a, 2011) argues that sociolinguists and typologists are actually referring to different contact situations: the former focus on those in which there is extensive short-term non-native acquisition and use by adult or adolescent learners, while the latter concentrate on cases of long-term contact within the same territory involving childhood bilingualism. The first type of situations produces simplification “due to the relative inability of adult humans to learn new languages perfectly” (Trudgill 2009a, 99), whereas the second type induces the appearance of new grammatical features due to additive borrowing, that is, the transfer of new features from one language to another which do not replace pre-existing ones (Trudgill 2011, 27), or due to contact-induced grammaticalization, a process whereby a new grammatical category emerges in a language on the model of another (Heine and Kuteva 2005, 7). Finally, low-contact situations involving the uninterrupted transmission of a language from one generation to the next not only promote the preservation of grammatical complexity but also foster the spontaneous proliferation of new grammatical categories, thus resulting in an increase in complexity. In line with these remarks, we need to distinguish between (at least) the following types of situations,

depending on the presence or absence of contact and, if present, on its intensity (cf. Trudgill 2009a, 2011):

- High-contact situations with short-term widespread second-language acquisition and use by adolescents and adults result in grammatical simplification due to imperfect learning.
- Situations involving intensive contact over a long period of time and extensive childhood bilingualism lead to grammatical complexification due to additive borrowing and contact-induced grammaticalization.
- Low-contact isolated communities preserve or even increase complexity due to the spontaneous growth of grammatical categories, among other processes.

From this typology, Trudgill concludes that it was in the third type of communities that grammatical complexity arose in the first place. As he argues:

If widespread adult-only language contact is a mainly post-neolithic and indeed a mainly modern phenomenon associated with the last 2,000 years, and if the development of large, fluid communities is also a post-neolithic and indeed mainly modern phenomenon, then according to this thesis the dominant standard modern languages in the world today are likely to be seriously atypical of how languages have been for nearly all of human history. We have become so familiar with simplification in linguistic change [...] that it has been tempting to regard it as normal – as a diachronic universal. Maybe, however, it is complexification that is more normal. Or rather, we should say, was more normal (Trudgill 2009a, 109).

The typical human community in pre-neolithic times was one in which contact with other groups was rare, and this is exactly the type of situation which is hypothesized to induce the growth of grammatical complexity. If this is true, then modern languages are not representative of how languages were for most human history, in which complexification was probably the most common type of development. The idea that languages can become simpler by chance with the passage of time is a consequence of the fact that most linguistic research was conducted in Western societies with histories of migration and, therefore, extensive language contact, which means that the languages spoken in those societies have been subject to simplification processes (cf. McWhorter 2008, 169).

2.2.4 Additional social determinants of complexity

Contact is not the only social factor that causes differences in complexity between languages. There are other social determinants of grammatical complexity, which are highly correlated with language contact but have their own independent effects. These factors, postulated by Trudgill (2009a, 2011) on the basis of previous sociolinguistic research, are the denseness of the social network and the size of the speech community (cf. also Kusters 2003, 2008).⁹ Communities with dense social networks are more able to preserve the complexities of their languages, and even to create new ones. This is because there is a tighter control over the learning process, so that the transmission of the language to the next generations is much more ‘perfect’ than in communities with loose ties. Additionally, small communities are able to preserve and also enforce grammatical features that increase complexity, because, for instance, in this type of communities an influential (group of) individual(s) can very easily

⁹ Further factors have been identified, such as the amount of communally shared information (Trudgill 2011, 127), or the use of a language for symbolic reasons (e.g. the expression of communal identity or some other aesthetic or ritualistic motivations) (Kusters 2003, 7).

compel the rest to use a new complex pattern. The denseness of social networks and community size are highly correlated with degree of language contact: high-contact communities are bound to be large in size and have loose social networks. However, these factors should be treated independently because, in theory, there could be, for example, a large community without a history of language contact, or a high-contact one with a dense social network.

2.3 MEASURING COMPLEXITY

Any discussion of grammatical complexity must necessarily be preceded by an explanation of how it is going to be measured, because there exists no agreement yet as to what the optimal way is to quantify the complexity of languages. Many metrics have been proposed, but they approach the study of complexity from different perspectives and with different research goals. In what follows, three different metrics postulated in the specialized literature will be described in detail and compared with the aim of extracting some useful generalizations that will be applied in Chapter 3 to measure the complexity of pronoun omission in English, the syntactic phenomenon examined in the present book. These metrics are the ones postulated by Kusters (2003, 2008), Miestamo (2006a, 2008), and Hawkins (1994, 2004, 2009, 2014).

2.3.1 Kusters (2003, 2008)

Kusters' (2003, 2008) metric was originally devised to assess the influence of social factors, such as language contact, on the complexity of verbal inflection. Kusters' approach departs from what he terms "a generalized outsider" (2008, 9), that is, an ideal user who has a mother tongue but does not speak the language to be measured and does not have any knowledge of the traditions and practices of the speech community in question. The main goal of this

speaker is to achieve a degree of competence in the second language that allows her or him to use it for communicative purposes but not necessarily for other types of practices related to more aesthetic or ritualistic ends.¹⁰ Complexity is then measured in relation to the preferences of this generalized outsider: those features that are difficult to acquire or process for this ideal speaker are categorized as complex, while those that do not require a great effort on her or his part are considered simple. More specifically, grammatical features that are easy to acquire for second-language learners are thus simpler under this definition, and those that pose problems for them are more complex. Additionally, a generalized outsider will judge as being more difficult (and, therefore, more complex in Kusters' metric) phenomena that are costlier to perceive or comprehend. In contrast, those that hinder production will not cause as many difficulties for this speaker, since there is always the possibility of modifying them to produce an imperfect but easier form. Finally, since the goal of a generalized outsider is to use the language for day-to-day communication, features that are less useful for this end and more adapted to symbolic uses are thus categorized as more complex.

Kusters then proposes three principles, violations of which increase the complexity of the morphological component of a language, and, more specifically, that of verbal inflection. These principles are the following:

- Economy (Kusters 2003, 22–25, 2008, 13), that is, the number of categories that are coded by means of inflectional affixes in the verb: the fewer categories, the more economical the language. For instance, a language which contains inflections for

¹⁰ A generalized outsider is an abstraction over real language users, each with a different first language. This avoids the problem of taking into account the influence of the mother tongue on the second language (cf. Kusters 2008, 9).

aspect and voice is simpler than one which gives overt morphological marking to aspect, voice, and agreement with the subject.

- Transparency (Kusters 2003, 26–30, 2008, 13), which refers to the clear coding of the categories: deviations from this principle by means of fusion (one form coding several categories syntagmatically), homonymy (one form for several categories paradigmatically), fission (many forms for one category syntagmatically), and allomorphy (many forms for one category paradigmatically) increase complexity.
- Isomorphy (Kusters 2003, 30–34, 2008, 13), that is, the order of the inflectional endings after the verb: the more similar the order is to a universal ordering of inflectional elements and, within the same language, to the order of elements in other semantic or syntactic domains, the higher the degree of isomorphy. The ideal isomorphic order, according to Kusters (2003, 32), is: verb stem > valency > voice > number object agreement > person object agreement > gender object agreement > aspect > tense > mood > number subject agreement > person subject agreement > gender subject agreement. Deviations from this ideal order and inconsistencies in the ordering of elements within the same language are considered to increase complexity by reducing its isomorphy.

These three principles are designed to quantify the complexity of verbal inflections, but at least the first two could be easily generalized to encompass other grammatical domains. Economy, for instance, could be extended to account for the number of categories encoded by any other morphological or even syntactic means. In a similar manner, transparency could refer to the clear coding of grammatical categories in any domain besides that of verbal inflection. Isomorphy seems at first sight to be less generalizable than the previous principles,

although its scope could be broadened to account for the consistency of the ordering of elements throughout all the different components of a grammar.

2.3.2 Miestamo (2006a, 2008)

Miestamo's (2006a, 2008) metric was designed for typological research and thus measures complexity within functional domains of grammar, such as the encoding of tense, negation, or core argument marking (cf. Footnote 4). Miestamo understands complexity in an arguably objective manner as the number of parts in a system, a definition that is reminiscent of that of Rescher (1998) mentioned above: the more parts a system has, the more complex it is. In order to operationalize this definition, Miestamo makes use of an idea, originally developed within the framework of information theory, which argues that the complexity of an object can be quantified on the basis of the length of the description that is necessary to account for this object (cf. Dahl 2004, 21). If we apply this to the study of functional domains in grammar, a language in which the encoding of a specific semantic/pragmatic category requires a longer description in comparison with that of another language is more complex. On this basis, Miestamo (2006a, 349–50, 2008, 33) postulates two criteria to measure complexity, the principles of Fewer Distinctions and One-Meaning–One-Form.

Fewer Distinctions stipulates that a language in which a larger number of semantic/pragmatic distinctions is grammaticalized in comparison with another entails a longer description and is thus more complex. For example, if language A distinguishes between present, past, and future within the domain of tense, and language B distinguishes only between present and past, then language A is more complex than language B. The One-Meaning–One-Form principle pays attention not to the number of distinctions within a particular functional domain, that is, grammatical meanings, as in the case of Fewer

Distinctions, but to the formal marking which is given to the meanings that are distinguished in that domain. According to this criterion, complexity increases if the coding of grammatical categories does not follow a one-to-one distribution – one category coded by one form – because this entails a longer description than a situation in which the mapping of meanings to forms is transparent.

Miestamo's (2006a, 2008) Fewer Distinctions and One-Meaning–One-Form principles clearly overlap with Kusters' (2003, 2008) Economy and Transparency in that, if applied to the same grammatical domain in the same languages, in most cases they will provide comparable assessments of their complexity. However, as will be explained in Section 2.3.4, they are diametrically opposed with respect to their conception of grammatical complexity and the perspective from which they approach its quantification.

2.3.3 Hawkins (1994, 2004, 2009, 2014)

Hawkins embeds the discussion of complexity within the more comprehensive concept of efficiency. In his metric, complexity (of a particular linguistic structure, not of a functional domain) increases when more forms and their associated properties must be processed to access the ultimate semantic and syntactic representation of a sentence. The term *form* is broadly construed to include phonemes, morphemes, words, and larger units such as phrases. *Property*, on the other hand, refers to lexical meanings, part-of-speech categories, semantic and syntactic requirements of co-occurrence, theta-role assignments, non-compositional meanings that depend on more than one form, and, more generally, to the construction of higher syntactic units such as noun or verb phrases, which are projected from the form functioning as the head of the phrase, as well as other syntactic relations like sisterhood (cf. Hawkins 2004, 15–18). Additionally, complexity increases if larger domains for the

assignment of properties to forms have to be established in on-line processing, and if the ordering of forms within the sentence delays the access to its ultimate semantic and syntactic representation. Efficiency is associated with the essential communicative function of language of transmitting information. According to Hawkins (2009, 253), efficiency increases when the speaker is able to relay a message to the addressee with the minimum processing effort required and within the shortest possible time frame. This entails that, to achieve an ideal level of efficiency, the speaker may sometimes have to resort to more complex structures and sometimes to simpler ones, depending on the minimum amount of complexity that is necessary to transmit the message that she or he wants to convey to the addressee in a particular situation.

Hawkins substantiates his definition of complexity by means of three more specific principles described as processing preferences of language users, violations of which increase the complexity of particular structures. These principles are not to be confused with those in Filipović and Hawkins (2013) explained in Section 2.2.2. They all derive from speakers' preferences with respect to complexity and efficiency but, whereas those in Filipović and Hawkins (2013) are based on second-language learners, the ones presented here are more general processing considerations.

- Minimize Domains (MiD): “The human processor prefers to minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed” (Hawkins 2004, 31). Combination refers to two or more forms that belong to the same phrasal unit and/or lexical co-occurrence frame, while dependency concerns a relation between forms whereby one relies on the other for the assignment

of a semantic or syntactic property (Hawkins 2004, 18–22). What MiD suggests is, then, that language users favour structures in which combination and dependency relations can be established without having to parse a large number of forms and their properties. For instance, if form A depends on form B, then, according to MiD, they should be positioned as close to each other as (grammatically) possible so as not to have to deal with extra material in the process of establishing such a dependency.

- **Minimize Forms (MiF):** “The human processor prefers to minimize the formal complexity of each linguistic form F (its phoneme, morpheme, word, or phrasal units) and the number of forms with unique conventionalized property assignments” (Hawkins 2004, 38). The idea is that processing formal units and semantic/syntactic properties involves an effort for language users, so they prefer to process as fewer forms and properties as possible. MiF implies four more specific predictions. First, there is a preference for fewer units per form, for instance, words with fewer phonemes and morphemes, phrases with fewer words, or sentences with fewer phrases. Second, minimizing the number of grammaticalized categories in a language is beneficial according to MiF, since this entails fewer forms and properties that have to be obligatorily produced and parsed in on-line processing. Third, more properties should be assigned to fewer forms by means of leaving some forms ambiguously or vaguely specified as to the property that must be assigned to them in a particular context, and by giving zero marking to some grammatical distinctions. Finally, the use of contextual information should be maximized in order to assign the correct property to ambiguous, vague, or zero forms. Contextual information in Hawkins’ theory includes, among others, high accessibility in the current discourse, high assignment

frequency of a property to a form in previous discourses, default assignments, and implicatures (Hawkins 2004, 44–45).

- Maximize On-line Processing (MaOP): “The human processor prefers to maximize the set of properties that are assignable to each item X as X is processed” (Hawkins 2004, 51). MaOP describes the preference of language users for structures that provide access to as much of the final semantic and syntactic representation of a sentence as early as (grammatically) possible, which implies being able to ascribe all the properties that must be assigned to a form as this form is processed and not later. If a form must be kept in memory until another form is reached to assign all its properties within a given context, processing effort (and, thus, complexity) increases. For example, antecedents do not tend to occur after the anaphoric element that refers to them (neither within a sentence nor in a larger stretch of discourse), because this would mean that the referent of the anaphor could not be correctly identified until the antecedent was parsed, which would make the user incur in higher processing costs.

As in other theories which propose multiple interacting factors (cf., for instance, Filipović and Hawkins 2013; Section 2.2.2), the complexity principles in Hawkins’ metric sometimes have a reinforcing and sometimes a contradicting effect. To illustrate this point, consider a hypothetical language with explicit coding of case, that is, with a set of forms, such as affixes, that overtly signal if a noun phrase is nominative, accusative, or dative within a particular construction. Overt case marking is disfavoured by MiF because it usually entails the production and parsing of more formal units, and thus higher processing costs. On the other hand, it is advantageous according to MaOP, since theta-role assignment can be dealt with earlier, as each noun phrase is processed, without having to delay it until the verb is reached.

Overt case marking is also beneficial with respect to MiD, since theta-roles can be processed as each noun phrase is encountered, which in turn means that the language user does not have to establish a dependency between noun phrases and the verb so as to assign their case.

Interestingly for the purposes of the present monograph, later formulations of Hawkins' theory make explicit reference to the interaction of efficiency and complexity with social factors such as language contact (cf. Hawkins 2014, 87–89). He argues that efficiency is useful to understand the temporality and directionality of changes that are caused by external variables. As explained above, in many cases the principles in Hawkins' metric contradict one another. This creates internal tensions within the grammar that can be exploited by certain groups of speakers, such as second-language users in the case that, at a certain point in the history of a language, they become a sizeable group within a speech community and leave their imprint in the grammar. Different types of speakers may have different processing preferences that can tilt the balance in favour of a particular linguistic variant and result in a change in usage frequency, which may eventually become grammaticalized (cf. Hawkins 2004, 3).

2.3.4 Comparison between the three metrics

The three metrics described in Sections 2.3.1, 2.3.2, and 2.3.3, are all attempts to quantify grammatical complexity. Therefore, they have some similarities but also many differences, particularly with respect to their conception of complexity, the objects they measure, the number of factors they propose, and the motivations underlying them. Kusters' and Miestamo's metrics postulate similar complexity principles: Kusters' Economy and Miestamo's Fewer Distinctions basically measure the same type of complexities and reach identical conclusions, that is, the larger the number of grammatical categories distinguished,

the more complex the language. Likewise, the Transparency and the One-Meaning–One-Form principles favour the clear coding of categories, so that one grammatical meaning should always be encoded by only one form, both paradigmatically and syntagmatically speaking.¹¹ They are also similar with respect to the types of objects that they measure. Both metrics include paradigms and structures in their assessment of complexity within the Transparency and One-Meaning–One-Form principles, respectively: the clear coding of grammatical categories should be complied with both syntagmatically (for instance, plurality should be coded only by one formal unit in a particular structure) and paradigmatically (for example, within a paradigm there should be only one distinctive form marking plural number). However, the definition of complexity is very different in Kusters’ and Miestamo’s metrics. On the one hand, Kusters sets off from the concept of a generalized outsider, and ideal second-language speaker whose preferences lie at the basis of his principles: what is difficult or costly for this type of speaker is complex according to his metric, that is, it is “user-oriented” (Miestamo 2006a, 346). This includes features that are difficult to learn, to perceive, and that are mostly adapted for symbolic uses. In Miestamo’s metric, on the other hand, the point of departure is description length: what requires a longer description is more complex. Crucially, the description of grammatical phenomena depends on the theory chosen by the analyst, so it can be said that Miestamo’s metric is “theory-oriented” (Miestamo 2006a, 346).

Kusters’ metric differs from that of Hawkins in the complexity principles they postulate. Kusters’ Economy principle and Hawkins’ MiF are principles of least effort, but they differ in that the former is basically restricted to the number of categories coded in the verbal inflectional paradigm (or any other domain), while the latter is much more general in that it

¹¹ Kusters’ Isomorphy principle does not have a counterpart in Miestamo’s metric since the order of elements does not occupy a prominent position in the latter’s theory.

can be applied to several types of phenomena and it makes a wider range of predictions: fewer units within a form, fewer grammaticalized distinctions, less clearly specified forms, and more reliance on contextual information for disambiguation (cf. Section 2.3.3). Isomorphy and MaOP are comparable in that they both refer to the ordering of units. However, their motivations, as well as the complexity measurements they provide, are quite different, the former focusing on the consistency of the arrangement of elements throughout the grammar and with respect to some universal order, and the latter favouring an order which provides an early access to the ultimate semantic/syntactic structure of a sentence. Finally, MiD does not have a counterpart in Kusters' metric, since the processing of syntactic and semantic combinations of and dependencies between elements seems not to play a role. Moreover, we do not encounter any explicit Transparency principle in Hawkins' metric, which only makes reference to this issue when discussing the limits of formal minimization: forms can be minimized only when contextual information (broadly construed; cf. Section 2.3.3) ensures the unambiguous mapping of properties to forms. A further, and important, difference between Kusters' and Hawkins' metrics is related to the types of objects they measure: whereas Kusters' metric, as discussed above, takes into account paradigmatic and syntagmatic violations of his Transparency principle, Hawkins' focuses exclusively on the complexity of specific structures. There is a minor reference in MiF to the number of conventionalized grammatical distinctions, but only because more conventionalized categories entail a larger number of forms and properties that must be obligatorily produced and parsed. On the other hand, these two metrics coincide in their definition of complexity, since both make reference to the preferences of language users, Kusters to those of his ideal second-language learner and

Hawkins to more general processing preferences, and what counts as complex is what is disfavoured by them.¹²

Finally, Hawkins' metric differs with that of Miestamo not only with respect to the principles they propose and the objects they measure (similar to the differences it has with Kusters' approach; cf. above), but also as regards their conception of complexity. The reader may recall from the previous discussion that Miestamo's definition of complexity was theory-oriented, that is, it measured complexity in terms of the length of the description of a linguistic phenomenon provided by a particular theory of grammar, while Hawkins claimed that what counts as complex are all those features that make the processing of linguistic structures costlier for language users.

It is worth noting that, even though they operationalize complexity in different ways, the three metrics share some common ideas, and we can in fact extract two general complexity notions from them. For lack of a better terminology, we will call them Economy and Clarity.¹³ Economy increases with fewer forms at any level of analysis (that is, morphemes with fewer phonemes, words with fewer phonemes and morphemes, phrases with fewer words, and sentences with fewer phrases) and conventionalized meanings, both paradigmatically and syntagmatically, and with fewer lexical, morphological, semantic, syntactic, or pragmatic rules mapping meanings to forms. This reflects both the preferences of language users, who favour structures and systems involving the least possible effort to learn and use, and the idea that complexity increases with description length, since, other things being equal, a language

¹² A minor difference in this respect between Kusters' and Hawkins' metrics is that the former accounts for both learning and usage costs, while the latter focuses only on processing (usage) considerations. This is due to the fact that Hawkins' theory lacks a consistent systemic approach to complexity.

¹³ These general and abstract notions should not be confused with Kusters' (2003, 2008) well-defined and operationalized principles of Economy and Transparency.

with fewer forms, conventionalized meanings, and rules requires a shorter description. Clarity, on the other hand, works in the opposite direction by limiting the amount of minimization that is possible in language. Clarity increases when there is an iconic one-to-one relation between meanings and forms both paradigmatically and syntagmatically but, and this is an important distinction with Kusters' (2003, 2008) and Miestamo's (2006a, 2008) metrics, taking into account the role of contextual information, broadly construed as in Hawkins (2004, 44–45), so as to allow, in certain cases, for departures from an ideal level of transparency. These two general notions underlie most principles and metrics of complexity postulated to date, including the ones reviewed in this section, and will be taken into account for the discussion of the complexity of pronoun omission in Chapter 3, Section 3.4.

2.3.5 Some important distinctions

Besides the general complexity notions introduced in the previous section, there are some further distinctions which are related to the perspective that analysts take when approaching complexity. The first is that between global and local complexity. Global complexity refers to the total amount of complexity in a language. We have already examined several theoretical, methodological, and empirical problems related to this approach in Section 2.1, so it will not be discussed further here. What most scholars do is to concentrate on local complexity, that is, on measuring the complexity of specific grammatical components (phonology, morphology, semantics or lexis, syntax, and pragmatics; cf. Szmrecsanyi and Kortmann 2012, 9, and references therein) or linguistic structures. This approach is feasible and provides interesting and valuable results, so it will be the one followed in the present work.

Another important distinction is that between absolute and relative complexity, postulated by Miestamo (2006a, 2008), which was implicit in the terms “theory-oriented” and “user-

oriented” employed in Section 2.3.4. The absolute approach defines complexity in an arguably objective manner as the number of parts in a system or the minimum length of description of a phenomenon. It does not rely on concepts such as ‘difficulty’ or ‘cost’ because it is not based on the preferences of language users (Miestamo 2008, 24) and, therefore, what is categorized as complex in this approach may not coincide with what they perceive as being difficult. The relative approach, on the other hand, explicitly describes complexity as the difficulty experienced by language users in order to learn or use a language (Miestamo 2008, 25), so complex means costly for language users. Both approaches have been criticized for several reasons. Kusters (2008, 8) argues that the absolute approach depends on the theory chosen by the analyst, because different theories provide different descriptions of grammatical phenomena. This crucially affects description length, and thus any assessment of complexity derived from it, so the absolute approach may not be so absolute and objective after all. Additionally, following generativism, most theories of grammar hypothesize that what they describe is some system of linguistic knowledge located in the mind of language users, so any description of grammatical phenomena based on a specific theory ultimately depicts the preferences of speakers (Kusters 2008, 5). Miestamo (2008, 26) criticizes the relative approach because of its dependency on a specific type of language users. Different types of users have different preferences. For instance, a grammatical feature that facilitates acquisition for first language speakers may be difficult for second-language learners. Additionally, assessing the difficulty or cost of processing a particular feature in a specific language can only be done on the basis of psycholinguistic evidence, which is not always readily available. The relative approach has an important advantage over the absolute one: it explicitly provides an explanation for complexity measurements in terms of language user preferences, while the absolute approach does not. It

is true that the relative approach has some limitations, but I believe that they are not as problematic as those of absolute metrics. For instance, by specifying the preferences of the different types of users accounted for in a particular piece of research, any potential problem of conflicting user tendencies is solved. Moreover, several theories of language processing and acquisition postulated on the basis of empirical evidence are now available and they can be resorted to for the assessment of complexity in terms of processing or learning difficulty (cf. Hawkins 1994, 2004, 2009, 2014).

A third distinction that is not usually explicitly mentioned is that between system and structural complexity (Dahl 2004, 42–44). System complexity refers to the complexity of the rules mapping meanings to forms, while structural complexity measures the objects that result from the application of those rules. Furthermore, within system complexity, Dahl (2004, 49) distinguishes between what he calls *tectogrammatics* and *phenogrammatics*: whereas tectogrammatics refers to the existence of different structural choices in a grammar (the more, the more complex), phenogrammatics concerns the rules that constraint the use of those options. For example, a past tense form of a regular verb in English, such as *watched*, contains two morphemes and is thus structurally more complex than its present tense form *watch*, which only contains one. Similarly, the possibility of forming interrogative sentences like *Do you like pie?* increases the system complexity of English, from both a tectogrammatical and a phenogrammatical point of view. This is because it implies the existence of an extra structural choice in the grammar (that is, tectogrammatics) and because the formation of interrogative sentences involves a set of rules, such as inversion, that are not necessary for generating declarative sentences like *You like pie* (phenogrammatics). System and structural complexity were indirectly mentioned in the discussion of Kusters', Miestamo's, and Hawkins' metrics in Sections 2.3.4. The reader may recall that Hawkins'

metric focuses only on the complexity of linguistic structures and, therefore, it exclusively measures structural complexity. On the other hand, both Kusters' and Miestamo's metrics had a systemic and a structural component: syntagmatic violations of Kusters' Transparency and Miestamo's One-Meaning–One-Form principles pertain to the domain of structural complexity, while paradigmatic violations relate to the assessment of system complexity (particularly to tectogrammatics, that is, the number of structural choices within a paradigm, but also to phenogrammatics as a larger number of choices tend to imply a larger number of formation rules). In the present monograph, system complexity and structural complexity will be used as proxies for acquisition and processing difficulty, respectively, the assumptions being that a more complex system is more difficult to learn, and more complex structures are more difficult to use, that is, to produce and comprehend.

Finally, we should also distinguish, in the parlance of Bisang (2009), between overt and hidden complexity. Overt complexity refers to the set of morphosyntactic structures used to code grammatical categories and rules. However, languages often leave unexpressed certain information that can be retrieved from context. The term hidden complexity is used to refer to the processes of pragmatic inference that addressees must undergo in order to be able to retrieve unexpressed information and successfully decode the messages conveyed by speakers. Bisang (2009, 35) identifies two different types of hidden complexity: (i) ambiguous grammatical markers which can fulfill more than one function, and (ii) vague structures which can correspond to more than one construction. In languages that display vague and/or ambiguous forms, addressees need to work harder to interpret the meaning of utterances, thus incurring in additional processing costs. Hidden complexity poses a problem for studies that focus exclusively on the overt marking of grammatical meanings leaving aside pragmatic inference, as they do not measure complexity as a whole. The present monograph

deals mainly with overt complexity, but it also considers hidden complexity when discussing pronoun omission in varieties of English (cf. Chapter 3, Section 3.4).

2.4 COMPLEXITY AND VARIETIES OF ENGLISH

Many scholars have extended the study of complexity to the field of variation in English, in many cases with the aim of “understanding the comparatively simple (i.e. language-internal complexity variation) before approaching the comparatively complicated (i.e. cross-linguistic complexity variation)” (Szmrecsanyi and Kortmann 2012, 14). One of these scholars is Peter Trudgill (cf., for instance, 2009a, 2009b), who proposes a division between high-contact and low-contact varieties of English. The former category consists of pidgins, creoles, second-language varieties, shift varieties, and dialect-contact (including standard) varieties, while the latter comprises mainly traditional dialects of English, such as those spoken in England (in areas located far from London), southern and eastern Scotland, Northern Ireland, Newfoundland, and the Appalachians.¹⁴ High-contact varieties exhibit the typical effects of having undergone contact with other languages and dialects, that is, grammatical simplification due to adult non-native learning and use. Low-contact varieties, on the other hand, show a tendency to contain more complex grammatical features (Trudgill 2009a, 104–9), in line with the typology of contact situations introduced in Section 2.2.3.

Following Trudgill’s “true typological split” (2009b, 315) between high- and low-contact varieties, Kortmann and Szmrecsanyi conducted a series of studies on complexity differences

¹⁴ Second language varieties are spoken in speech communities with a majority of non-native speakers but play an important role in society by virtue of being the language used in administration, education, politics, etc. Shift varieties include those which were mostly second language varieties in the past but have now become the first language for the majority of the speech community. Dialect-contact varieties are the result of the convergence of different dialects of English in the same territory.

between dialects of English (cf. Szmrecsanyi and Kortmann 2012, 14–18, for a summary of their findings). Their first line of research analyses the survey data included in the *Handbook of Varieties of English* (Kortmann et al. 2004), which contains information about the attestation of seventy-six morphosyntactic features in forty-six varieties of English. The features are classified as to whether they increase or decrease the complexity of a variety on the basis of three specific measures (Kortmann and Szmrecsanyi 2009, 272–74; Szmrecsanyi and Kortmann 2009a, 64–65):

- Ornamental rule complexity: those features that increase the number of grammatical categories or distinctions beyond what is necessary for communication.
- Rule simplicity: those that decrease the number of categories or distinctions in the system.
- Second-language acquisition difficulty: features that occur frequently in second-language learner production.

Low-contact L1s, that is, traditional dialects of English, attest more ornamental rules than high-contact L1s (including transplanted and shift L1s; cf. Szmrecsanyi 2009), L2s and pidgins and creoles, while for the most part the opposite is true for rule simplicity and second-language acquisition difficulty: pidgins and creoles attest more of these features than high- and low-contact L1s and, surprisingly, L2s. A clear hierarchy emerges by classifying the features as to whether they increase or decrease a variety's morphosyntactic complexity and degree of analyticity (Szmrecsanyi and Kortmann 2009b, 1655–57): pidgins and creoles are morphologically less complex and more analytical than L2 varieties, which in turn are less

complex and more analytical than high- and low-contact L1s.¹⁵ These results point to the conclusion that high-contact varieties of English are in general simpler than low-contact ones, and within the former type, pidgins and creoles appear to be the simplest (cf. Section 2.2.1).

Kortmann and Szmrecsanyi's second line of research compares the complexity of varieties by means of four frequency indexes drawn from corpus data (Kortmann and Szmrecsanyi 2009, 2011; Szmrecsanyi 2009; Szmrecsanyi and Kortmann 2009a). The first index measures the incidence of analytical grammatical markers in one thousand decontextualized tokens of running text, while the second measures the frequency of synthetic grammatical markers (inflections). The third index results from the sum of the analyticity and syntheticity indexes, thus quantifying the occurrence of any type of grammatical markers. Finally, an irregularity index assesses the rate of irregular synthetic markers. Their findings indicate that whereas low-contact L1s score high in both syntheticity and analyticity, L2s are neither extremely synthetic nor analytic, with high-contact L1s occupying an intermediate position. This means that the crucial index is the third one, which measures a variety's grammaticity, that is, the incidence of both synthetic and analytic grammatical markers, with high-contact varieties favouring fewer explicit grammatical markers overall. Furthermore, low-contact L1s show a higher rate of irregular markers than high-contact L1s, which in turn are more irregular than L2s. These results are comparable to those of Kortmann and Szmrecsanyi's first line of research and suggest, again, that complexity in English varies as a function of a variety's degree of contact, with high-contact ones exhibiting less complexity.

Williams (1987, 162) concentrates exclusively on "non-native institutionalized varieties of English" or NIVEs, that is, second-language varieties with an important role in education

¹⁵ Analyticity is assumed to increase transparency and thus ease comprehension, which means that it decreases complexity (Szmrecsanyi 2009, 322).

and other functional and/or sociolinguistic domains, such as Indian English or Ghanaian English. She argues that these varieties share several commonalities that cannot be explained on the basis of nativization processes or as a function of first language influence, since they occur in many varieties located in different geographical areas. Additionally, these shared features can be found in the production of second-language learners, so they are better explained on the basis of acquisitional and processing mechanisms. First, there are certain structures or grammatical domains which pose problems for language learners, who may find them difficult to acquire or use, and, therefore, these features are commonly modified in the process of second-language acquisition. Second, these modifications tend to derive from processing preferences that are evident in learner speech.

Two production and comprehension principles are postulated by Williams (1987, 169–91): the principle of economy and the principle of hyperclarity. Economy describes the preference of speakers for efficient economical production, which minimizes the effort that they have to make to convey a message to the addressee. It accounts for many characteristic features of NIVEs, such as the omission of inflectional endings, pronouns, and copulas, as well as regularization processes like the use of invariant question tags or the extension of the progressive aspect to stative verbs. The principle of hyperclarity, on the other hand, illustrates the hearer's preference for unambiguous structures by means of transparent marking of grammatical categories, such as the occurrence of resumptive pronouns in relative clauses. Transparency also increases as a result of maximizing saliency, sometimes through the use of redundant markers as in the case of double conjunctions (e.g. *Though the farmer works hard, but he cannot produce enough*; Nihalani, Tongue, and Hosali 1977). The principles of economy and hyperclarity influence the production of second-language learners and NIVE speakers to generate a maximally economical and clear output. Williams' findings, therefore,

point to the conclusion that NIVEs are simpler varieties, in comparison with native ones, because they tend to contain grammatical features that increase economy and clarity, in line with the general complexity notions postulated in Section 2.3.4.

Davydova (2011) focuses also on non-native varieties of English, including, on the one hand, Indian English, East African English, and Singapore English, second-language institutionalized varieties in the sense of Williams (1987), and, on the other, foreign-language varieties like the English spoken in Russia and Germany. Taking as a point of departure the functional domain of tense, and more specifically the present perfect, she proposes two complexity metrics; one absolute and one relative. The absolute metric measures the complexity of present perfect marking as a function of the “degree of language-internal variation within a circumscribed variable context” (Davydova 2011, 103). This encompasses, first, the number of variants within the context, and, second, the number of constraints regulating the choice between the variants, a definition that is reminiscent of Dahl’s (2004) distinction between tectogrammatical and phenogrammatical system complexity explained in Section 2.3.5 (cf. also Huber 2012). The relative metric measures the amount of effort on the part of a learner to acquire and use a grammatical feature. According to the first metric, then, a variety’s complexity increases if there are extra structural variants in present perfect contexts, such as the base form of the verb or the simple past tense form, instead of only the *have*-perfect, and if more factors constrain the variation between these different choices. With respect to the second metric, a variety is more complex if it exhibits a higher percentage of correct uses of the *have*-perfect, which is the most difficult English verb form for learners, as attested by the fact that it is the last one they acquire and master (cf. Davydova 2011, 108).

The results of Davydova’s study point to a trade-off between absolute and relative complexity. On the one hand, foreign- and second-language varieties, especially basilectal

and mesolectal ones, are more complex than (native) Standard English and acrolectal varieties, since the former exhibit more structural variants in present perfect contexts and more factors constraining their use. On the other hand, standard and acrolectal varieties show a higher percentage of correct uses of the *have*-perfect form in present perfect contexts, and thus are more complex in terms of learning difficulty, than foreign- and second-language basilectal and mesolectal varieties of English. These findings are very interesting in that they highlight the explanatory power of relative metrics of complexity. From an absolute point of view, complexity increases in non-native varieties as more variants are included in the system and thus more constraints are necessary to account for them. However, from a relative perspective, this increase in absolute complexity can be explained as a result of the learners' struggle to acquire and use a complex form such as the *have*-perfect, replacing it in many cases with simpler ones. I believe that these results emphasize the importance of considering the preferences of speakers in any account of grammatical complexity, since what matters seems to be what users experience as being difficult/easy for them and not some abstract measure of the complexity of a system.

It is worth mentioning two more studies on complexity in varieties of English, since they deal with syntactic complexity, an aspect that has not received the attention it deserves. Brunner (2014, 2017) is concerned with noun phrase complexity in British English, Singapore English, and Kenyan English, measured as the amount of modification and the number of embedded noun phrases within a noun phrase – the more modifiers and the more embedded noun phrases, the more complex. Brunner's results show that noun phrases in British English tend to be more heavily modified and to contain more embedded noun phrases than in Kenyan English, with Singapore English occupying an intermediate position. These findings can again be interpreted against the backdrop of language contact effects: high-contact varieties are

simpler, in this case in the noun phrase domain, than low-contact ones. Additionally, as argued by Brunner (2014, 43), the intermediate position of Singapore English is suggestive if compared to its degree of development in E. Schneider's (2003, 2007) Dynamic Model of the evolution of Postcolonial Englishes:¹⁶ Singapore English is in a more advanced phase than Kenyan English, with an increasingly large number of children speaking it as their mother tongue and thus a weaker impact of simplification processes due to second-language acquisition and use.

Finally, Suárez-Gómez (2017) focuses on relative clause complexity as a function of transparency, understood in terms of the one-meaning-one-form rule, in British English, Indian English, Hong Kong English, and Singapore English. In her metric, relative clauses are more complex if they exhibit zero relativizers (one meaning with no formal expression), animacy agreement by means of *wh*-pronouns (because agreement increases redundancy and thus complexity), and if they are not adjacent to the head noun phrase (because a meaning has to be mapped to a discontinuous formal unit). Her results point yet again to the conclusion that high-contact varieties are simpler than low-contact ones: the high-contact Asian varieties in Suárez-Gómez's study generally disprefer zero relativizers and *wh*-pronouns in favour of invariant and non-redundant *that*, with the exception of Indian English, which shows a high frequency of *wh*-pronouns due to substrate influence (Suárez-Gómez 2017, 223). Additionally, two further innovative structural choices that increase transparency vis-à-vis the standard variety are found in her data. One is the use of resumptive pronouns, which fill the position of the relativized noun phrase within the relative clause and thus make the mapping

¹⁶ Schneider's Dynamic Model seeks to explain the evolution of postcolonial varieties of English on the basis of both the interaction between the indigenous and settler populations in a territory to which English was taken by means of colonization processes, and subsequent changes in identity as a result of a further integration of these two groups (cf. E. Schneider 2003, 2007, chap. 3). This model, despite being highly influential in the field, will not figure prominently in the present monograph.

of meanings to forms clearer; the other is the deletion of stranded prepositions, which increases transparency by omitting an element with a fairly predictable or even non-functional meaning.

The studies reviewed in this section reach a similar conclusion with respect to the relation between complexity and contact in varieties of English. High-contact varieties, including pidgins, creoles, second-language varieties, shift first languages, and dialect-contact or transplanted first-language varieties, are simpler generally speaking than low-contact ones, that is, traditional dialects. They are simpler because: (i) they are less morphologically complex and more analytical than low-contact varieties; (ii) they exhibit a lower frequency of (synthetic and analytical) grammatical markers in running discourse, as well as fewer irregular synthetic markers; (iii) they contain features that make structures more economical from the perspective of the speaker but also more transparent for the addressee in comparison with the standard variety; (iv) they substitute more complex forms with simpler ones, as in the case of the *have*-perfect; and (v) they are syntactically simpler, at least with respect to the domains of noun phrase and relative clause formation. At the cross-linguistic level, languages with a history of contact in which adult non-native learning and use dominated were shown to have simpler grammars than those in which contact did not play an important role (cf. Sections 2.2.1 and 2.2.3). The evidence examined in the present section suggests that this tendency also seems to apply between dialects or varieties of the same language.

3 PRONOUN OMISSION

Pronoun omission is a linguistic phenomenon that has been widely researched in many different languages and from a broad variety of perspectives and theoretical frameworks; it has figured prominently in generativist research (e.g., Chomsky 1981; Radford 2004), pragmatics (e.g., Y. Huang 1992, 2000), cognitive linguistics (e.g., Ariel 1988, 1990, 1994, 2001), and sociolinguistics (e.g., Otheguy, Zentella, and Livert 2007; Nagy et al. 2011), among others. Such a diversity of approaches to the same phenomenon has logically resulted in a variety of labels that have been used to refer to the same or similar mechanism, like pro-drop, zero anaphora, and pronoun omission/deletion. Pro-drop is the term used in formal analyses of pronoun omission and assumes the existence of an empty element in the clause with a function from the point of view of syntax but no overt phonetic realization (Haegeman 1995, 451). Zero anaphora entails that the omitted element has an antecedent that must be recovered from the preceding discourse, thus excluding cases of omission of non-referential pronouns. Pronoun omission/deletion will be the label used in the present book because it does not imply the existence of empty categories nor does it exclude non-referential deleted elements, which will be included in the cross-varietal study which is the focus of Chapter 4.

As a first approximation to the topic at hand, the following definition of pronoun omission can be provided: pronoun omission refers to the existence of a gap in a clause that could be filled by an overt element, usually (but not exclusively) a personal pronoun. This is a

maximally general definition that does not exclude non-referential pronouns and does not specify a particular syntactic function or grammatical role for the omitted pronoun, that is, it includes omitted pronouns in subject or object (or any other) position. Additionally, it does not stipulate the existence of an empty element when there is not an overt one. The following examples, extracted from *The Electronic World Atlas of Varieties of English* (eWAVE; Kortmann and Lunkenheimer 2013), a database containing information on 235 morphosyntactic features of spontaneous speech mapped onto 76 varieties of English, illustrate the occurrence of referential and non-referential deleted pronouns in subject and object positions in several varieties:

- (2) **He_i** did two year there, was it or something like that, **Ø_i** went fae there to London, **Ø_i** met up with his wife there. (Example 4026; Scottish English)
- (3) **Ø** Is highly unlikely that he will be here on time! (Example 3440; Pakistani English)
- (4) They don't give us **allowances_i** at school. So the first day my dad gave **Ø_i** me but afterwards he wasn't giving **Ø_i** me. (Example 3027; Ghanaian English)
- (5) I like **Ø** when she says nice things like that. (Example 749; Sri Lankan English)

Example (2) is a case of omitted referential pronouns in subject position: the subjects of *went* and *met*, which refer back to the overt personal pronoun *He*, are deleted. In (3), we have an instance of an omitted non-referential subject pronoun namely *it*, which would, under normal circumstances, anticipate the following extraposed subject clause: *It is highly unlikely that he will be here on time*. (4) and (5), on the other hand, contain instances of omitted objects: in (4) there are two such deleted elements, which refer back to the noun phrase *allowances*, while in (5) the omitted object is non-referential, i.e. *I like it when...* etc.

In the rest of this chapter, an overview of the most important theoretical approaches to the omission of pronouns is presented, with a focus on the conditions identified in the literature that license the occurrence of these elements in several languages (Sections 3.1 and 3.2). Section 3.3 deals with pronoun omission in English (in both standard and non-standard varieties) and with the major factors that account in this language for the variation between deleted and overt pronouns. Finally, Section 3.4 provides an analysis of pronoun omission in terms of its complexity, on the basis of the concepts described in Chapter 2.

3.1 GENERATIVE APPROACHES

Generative analyses of pronoun omission focus on uncovering the syntactic conditions under which pronouns can remain unexpressed. Rizzi (1986), in one of the most important early contributions to the issue, argues that an omitted pronoun is subject to two conditions: (i) it must be licensed by a governing head, and (ii) its grammatical features must be recovered from this head. In rich agreement languages, agreement is the head that licenses the occurrence of omitted elements, with their content being recovered from the person, number, and gender features coded in the agreement affix. This is the reason why rich agreement languages tend to allow pronoun omission (Dryer 2013). Additionally, Rizzi points out that omitted objects in, for instance, Italian, must receive an arbitrary interpretation, since their features cannot be recovered from agreement affixes, there being no object verb agreement system in this language. Mandarin Chinese, Japanese, and Korean, among others, pose a problem for this proposal because they do not have an agreement system at all (neither with subjects nor with objects) but still allow the omission of pronouns. C.-T. J. Huang (1984) claims that in these languages omitted subjects are in most cases variables coindexed with an

empty topic, as illustrated in examples (6) and (8) from Mandarin Chinese.¹⁷ Alternatively, they can also refer back to the subject of the main clause if they occur in a subordinate clause, as in (7).

(6) [$\emptyset_i$] $\emptyset_i$ lai-le.¹⁸

[**Himi**] $\emptyset_i$ came.

(7) **Zhangsan**_i shuo [$\emptyset_i$ bu renshi Lisi].

Zhangsan_i said [$\emptyset_i$ did not know Lisi].

(8) [$\emptyset_i$] Zhangsan shuo [$\emptyset_i$ bu renshi Lisi].

[**Himi**] Zhangsan said [$\emptyset_i$ did not know Lisi].

The idea is that in (6), the subject, a person that is not mentioned in the sentence but has been mentioned previously in the discourse, first moves to a topic position and is then deleted, leaving behind a variable that is coindexed with this omitted sentence topic. In (7), on the other hand, the omitted subject is coindexed with the subject of the main clause *Zhangsan*, but it could also refer to a deleted sentence topic, that is, someone that is not *Zhangsan* but has been mentioned before in the discourse, as illustrated in (8).¹⁹ This ambiguity, however, is not

¹⁷ Generative grammar distinguishes between four different types of empty elements: NP-traces, *pro*, PRO and variables. Out of these, *pro* and PRO are generated in situ, while NP-traces and variables are left behind by movement operations (e.g. raising or topicalization) and are coindexed with the moved element. The antecedent of a variable moves to a nonthematic position, such as sentence topic (example (6)), while that of a NP-trace moves to a thematic one, such as subject or direct object (cf. Chomsky 1981, 1982).

¹⁸ Examples 6-9 are adapted from C.-T. J. Huang (1984).

¹⁹ The occurrence of empty topics is a property of discourse-oriented languages like Mandarin Chinese, and this is the reason why we encounter the ambiguity shown in examples (7) and (8) (C.-T. J. Huang 1984).

found in the case of omitted objects, as in example (9), where, according to C.-T. J. Huang, the deleted element can only be a variable coindexed with an empty sentence topic:

(9) [$\emptyset_i$] Zhangsan shuo [Lisi bu renshi $\emptyset_i$].

[**Him_i**] Zhangsan said [Lisi did not know $\emptyset_i$].

This situation is hypothesized to derive from C.-T. J. Huang's (1984, 552) "Generalized Control Rule" (GCR), which states that an omitted pronoun must be coindexed with the closest nominal element, including agreement affixes attached to the verb. In (6), there is not a noun phrase in the clause with which the omitted subject can be coindexed and no subject verb agreement, so C.-T. J. Huang argues that the only possibility is for it to be a variable coindexed with a deleted sentence topic. In (7), the main clause subject noun phrase can act as an antecedent for the omitted pronoun in the subordinate clause, or it can also be a variable coindexed with a deleted topic, as in (8). Finally, the omitted object in (9) can only be the result of topicalization and subsequent deletion of the sentence topic. This is due to the interaction of the GCR with Chomsky's (1981, 188) binding condition B, which states that a pronoun cannot refer to an element within its governing category, that is, the minimal domain (a noun phrase or a clause) that contains the pronoun and a governing head (the subject in example (9)). According to the GCR, the omitted object should be coindexed with the closest nominal element, that is, *Lisi*, but this is blocked by condition B. Additionally, the subject of the main clause, *Zhangsan*, cannot act as an antecedent for the omitted object because it is not the closest nominal element, meaning that this interpretation is not possible under the GCR. Therefore, the only option is for the deleted object to be a variable coindexed with an empty sentence topic. At the time, C.-T. J. Huang's theory seemed to provide a valid solution to the

fact that omitted elements occur both in languages with rich agreement and in those without any agreement at all. However, as pointed out by Y. Huang (2000, 85–86), the requirement that matrix subjects cannot be antecedents of omitted subjects in subordinate clauses is not always true, that is, depending on the context, the deleted element in, for instance, example (9), can also refer to *Zhangsan*.²⁰ Further problems with C.-T. J. Huang's analysis are identified by Xu and Langendoen (1985) and Xu (1986) on the basis of the subjacency and bijection principles, and strong crossover conditions in Chinese.

Jaeggli and Safir (1989) propose a different solution to account for the apparent contradiction that omitted elements are possible both in languages with rich agreement and in languages with no agreement. They argue that deleted subjects only occur in languages which are morphologically uniform, that is, which have only derived or underived inflectional forms. Derived forms are those that can be decomposed into a root plus an affix, with affix being broadly construed to include suffixes, prefixes, and infixes, but also other processes such as suppletion or reduplication. Underived forms, on the other hand, are those that cannot be distinguished from the stem. English, for instance, contains both derived (e.g., 3rd person singular present verbal inflection *-s*, as in *plays*) and underived forms (e.g., 1st/2nd person singular/plural and 3rd person plural *play*), so omitted subjects cannot occur in this language, according to Jaeggli and Safir. Speas (1994, 2006) explains this generalization on the basis of the inherent content of agreement affixes. She claims that, in rich agreement languages with only derived forms, each inflection has its own lexical entry and thus inherent specified morphological features. In poor agreement languages with some derived and some underived forms, on the other hand, affixes do not have independent entries in the lexicon, so they do not have morphological content. Therefore, poor agreement morphology must receive its

²⁰ This is also a problem for Rizzi's (1986) theory, since it shows that not all omitted objects are arbitrary.

features from the subject, which must inevitably be overt in order to do so. This is not necessary in rich agreement languages because inflections have content. In languages with no agreement altogether, subjects can also be omitted since there is no agreement morphology in the clause that depends on them for the specification of its features. However, according to Cole (2009, 2010), Swedish, Danish, and Afrikaans, among others, are morphologically uniform languages but do not allow subject omission.

More recent formal approaches (cf., for instance, Tomioka 2003; Frascarelli 2007; Holmberg, Nayudu, and Sheehan 2009), also fail to account for the whole range of data regarding omitted pronouns. Neeleman and Szendrői (2007) propose a binary split of languages in terms of the role of agreement: there are agreement-based languages, in which the deletion of pronouns depends on their agreement morphology, and there are others in which agreement does not play a role. In the former type, omitted pronouns occur if the agreement morphology is rich enough to identify them. In the latter, as argued by Neeleman and Szendrői (2007), deleted elements only occur if pronouns are marked agglutinatively for case or other morphological distinctions. Their theory accounts for pronoun omission in 20 languages, but there are also many counterexamples: Finnish, for instance, is marked agglutinatively for case but lacks omitted pronouns (Cole 2009).

In the face of this situation, that is, that no theory to date has been able to explain all the facts related to pronoun omission exclusively on the basis of syntactic restrictions, Cole (2009, 2010) claims that the syntactic licensing of deleted pronouns is not only unnecessary, but also inadequate. He argues that it is not clear why a phonological distinction, namely omitted versus overt, should be explained in terms of a syntactic mechanism. Additionally, other factors, such as the accessibility of referents in discourse in the case of referential

pronouns, must be obligatorily taken into account in order to reach a fuller understanding of this phenomenon.

3.2 COGNITIVE APPROACHES

Formal approaches to pronoun omission highlight the role of agreement in order to account for the occurrence of deleted pronouns. Whereas it is certainly true that this is a very important factor to consider, at least in those languages in which it does play a role, it cannot provide a full description of pronoun omission. On the one hand, even though there seems to be a connection between agreement and pronoun omission, in that many languages that exhibit one feature also exhibit the other, this is not always so: there are rich agreement languages, such as Icelandic (Cole 2009), that do not allow deleted pronouns. On the other hand, there are languages without agreement, for instance, Chinese, in which pronouns can remain unexpressed, and even more frequently than in rich agreement languages (Neeleman and Szendrői 2007).

Other approaches focus on aspects that go beyond the syntactic environment in which deleted pronouns can or cannot be found, and adopt instead a more cognitive perspective, thus taking into account issues such as the availability of referents in discourse, the role of frequency, or processing considerations (e.g., priming effects), while at the same time retaining agreement as a potential explanatory factor in those languages in which it occurs.²¹

²¹ There are also theories that seek to account for pronoun omission and other anaphoric processes in terms of pragmatic inferences. The main idea is that reduced, semantically general anaphoric elements select locally coreferential interpretations, while larger and more specific forms select locally non-coreferential ones. This is explained on the basis of a series of underlying pragmatic principles of conversation and implicature (cf. Y. Huang 2000, chap. 4). These approaches will not figure prominently in the present work, so they will not be further discussed.

3.2.1 Agreement and context

Cole (2009, 2010) proposes a theory which takes into account the role of both agreement morphology and the context in the occurrence of referential omitted subjects.²² He recognizes the importance of agreement on the basis of cross-linguistic correlations, because omitted pronouns tend to occur in rich agreement languages, but also in the face of language internal evidence. Pashto, a split ergative language spoken in Afghanistan and parts of Pakistan, is a particularly clear example of the importance of agreement in some languages (cf. C.-T. J. Huang 1984, 535–36, for a detailed account and examples). In the present tense, the subject and the verb agree in transitive and intransitive clauses, and thus the subject can be omitted. In the past tense, however, the verb agrees with the subject in intransitive clauses but with the object in transitive ones, and only those pronouns with which the verb agrees can be omitted, that is, the subject in the former case and the object in the latter. Similarly, in many other languages, such as Swahili, Georgian, and Arabic (C.-T. J. Huang 1984), pronouns can only be dropped when the verb agrees with them but not when it does not.

Agreement, however, despite being a crucial factor in some languages, cannot account for all the facts related to pronoun omission. Even in rich agreement languages with omitted pronouns, agreement is not always enough to retrieve the morphological features of the deleted element. As shown by Cole (2009, 567), Spanish, a rich agreement language with omitted subjects, has several verbal endings that are ambiguous, like *tenía* in example (10), which can be either a 1st or a 3rd person singular form:

(10) **Juan**_i llegaba a casa. **Ø**_i Tenía las llaves.

²² As argued by Cole (2009, 2010), his theory could also be extended to include referential omitted pronouns in other positions. Non-referential pronouns, on the other hand, require a different approach and may even be subject to syntactic licensing.

Juan_i was arriving home. **He_i** had the keys.

The verb in the second clause of example (10) is ambiguous between a 1st and a 3rd person interpretation. The subject can still remain unexpressed, however, because there is an accessible antecedent in the previous clause that provides the morphological features that cannot be retrieved from the verb. On the other hand, this is not the case in (11), where we have two potential antecedents for the subject of the second clause (*Juan* and *yo*). An overt pronoun is obligatory in this example unless the previous discourse helps disambiguate:

(11) **Juan_i** y **yo_j** llegábamos a casa. ***Ø_{i/j}**/**Él_i**/**Yo_j** tenía las llaves.

Juan_i and **I_j** were arriving home. **He_i**/**I_j** had the keys.

A similar situation holds in other rich agreement languages, such as European Portuguese, Turkish and Serbian (Cole 2009).

Languages such as Spanish suggest that we need to consider (at least) both agreement inflections and an antecedent in the context if we want to explain the occurrence of referential omitted pronouns. As argued by Cole (2009, 2010), a combination of both factors is necessary: a deleted subject may occur if agreement identifies its features up to the point of a language's *morphological maximality*, and if the rest of the features are retrieved from a referent in the previous discourse. Morphological maximality refers to “the maximum extent to which null subjects can be recovered by agreement in the language concerned” (Cole 2010, 279), that is, the largest set of features that can be unambiguously retrieved from the agreement affixes attached to the verb. In English, for instance, these would be person and number, but only in the present tense and in certain forms: in the 3rd person singular form of

most verbs (except modals), and the 1st and 3rd person singular forms of the verb *be* (*am* and *is*). In all other cases, either there is no subject verb agreement, or the forms of the verb are not enough to identify the person and number of the subject (e.g., *you/we/they were*), so reference must be made to an antecedent in the previous discourse. Additionally, gender is not coded in English verbal morphology, so it must necessarily be retrieved from the context. The morphological maximality of a language such as Chinese, where there is no agreement system at all, is effectively zero, because no features can be retrieved from the verbal morphology. In these languages the only source for the identification of the morphological features of omitted pronouns is the context.

In Cole's theory of pronoun omission, a subject can be left unexpressed if:

- there are agreement affixes in the verb that identify the morphological features of the omitted pronoun up to the language's morphological maximality, and
- the rest of the features can be retrieved from an antecedent in the context.
- If there is neither subject verb agreement nor an accessible antecedent, some languages can resort to a default interpretation (e.g., 1st person singular in Italian; cf. Cole 2009, 570).

If none of the above conditions hold, then an overt pronoun must be used. The role of agreement in Cole's theory is minimized to that of an aid to recover the morphological features of deleted elements. It is neither a necessary nor a sufficient condition, thus fitting the observable cross-linguistic data on the occurrence of omitted pronouns. The real difference between languages that allow the deletion of pronouns and those that do not lies elsewhere, and not on the richness of their verbal morphology. It is their contextual strength/weakness,

understood in terms of whether they permit the retrieval of morphological features from the context, which differentiates languages that omit pronouns from those that do not. Spanish is contextually strong, so the features that are not recovered from agreement inflections can be recovered by reference to an antecedent in context. Chinese is also contextually strong and, consequently, it allows the omission of pronouns even without agreement. Icelandic, on the other hand, is contextually weak: it does not allow the omission of pronouns despite exhibiting rich agreement. Similarly, in Swedish pronoun omission is not permitted because it is a contextually weak language without agreement morphology. The following classification of languages according to their contextual strength and the role that agreement plays in the retrieval of morphological features is then proposed (Cole 2010, 303):

- The first group comprises languages that are contextually weak; antecedents in these languages can (almost) never be referred to by means of omitted pronouns regardless of the richness of their verbal morphology.
- The second group consists of contextually strong languages in which agreement mostly determines the occurrence of omitted pronouns, with reference to the context in those cases in which agreement is ambiguous. Additionally, this category includes languages in which pronouns can only be omitted when agreement recovers all their morphological features.
- The third and final group is made up of contextually strong languages without any agreement altogether; these languages also allow deleted pronouns because morphological features can be retrieved from antecedents in the context.

Cole (2010) operationalizes the notion of contextual strength/weakness on the basis of Ariel's (1988, 1990, 1994, 2001) accessibility theory, which is the focus of the next section. This theory postulates that there is a correlation between anaphoric elements and the degree of mental accessibility which they code: omitted pronouns are markers of high accessibility and can only be used to refer to highly accessible antecedents. Therefore, in Cole's terms, contextually strong languages are those in which certain antecedents are regarded as being so highly accessible that even minimal forms, that is, omitted pronouns, can be used to refer to them. In contextually weak languages, on the other hand, no type of antecedents is considered to be accessible enough to be referred to by means of deleted pronouns.

3.2.2 Accessibility theory

Ariel's (1988, 1990, 1994, 2001) accessibility theory proposes that referring expressions code the degree of mental accessibility of the entities to which they refer and are thus used by the speaker/writer to direct the addressee to the correct interpretation. The speaker indicates which is, in her or his opinion, the degree of accessibility of a given entity or event in the addressee's memory by means of a more or less explicit anaphoric element. Each of these elements signals a specific level of accessibility on the basis of three highly correlated factors: informativity, rigidity, and attenuation. Informativity alludes to the amount of lexical information of the anaphoric expression: for instance, noun phrases are more informative than pronouns. Rigidity relates to how ambiguous in reference a form is; proper names are highly rigid, while pronouns are not rigid since they can refer to almost any entity or event. Finally, attenuation concerns the phonological size of the anaphoric element: omitted pronouns are highly attenuated, while noun phrases are less attenuated. The interaction between these three

factors results in the following hierarchy of anaphoric devices, organized with respect to the accessibility which each of them encodes (Ariel 2001, 31):

Full name + modifier > full name > long definite description > short definite description > last name > first name > distal demonstrative + modifier > proximate demonstrative + modifier > distal demonstrative + NP > proximate demonstrative + NP > distal demonstrative (- NP) > proximate demonstrative (- NP) > stressed pronoun + gesture > stressed pronoun > unstressed pronoun > cliticized pronoun > verbal person inflections > zero

The expressions at the beginning of the scale are low accessibility markers because they are highly informative, rigid and unattenuated, so they are used to refer to antecedents which are not very accessible. Those at the end, on the other hand, code a high degree of accessibility. Zero, the last item on the scale and the one referring to omitted pronouns, does not contain any kind of lexical information, it is not rigid in reference, and it is extremely attenuated, thus being used to refer only to highly accessible antecedents.

Ariel argues that we can measure the degree of accessibility of an entity or event on the basis of its salience in discourse and the strength of the connection between the anaphoric element and its antecedent. Several factors increase the salience of a referent. First, the participants in a conversation, that is, the speaker and the hearer, are more salient than any 3rd person referent that is not in the immediate physical context.²³ Second, topics are more salient than non-topical entities and, similarly, humans are typically more salient than non-humans. Third, highly accessible referents usually occur in subject position, so subjects are considered to be more salient than non-subjects. Furthermore, salience also varies as a function of the

²³ This notion could be extended to include the writer(s) and the reader(s) of a text.

ambiguity of an entity, that is, if there are other potential antecedents in the immediate discourse, the salience of a referent decreases. Finally, accessibility is increased by low accessibility markers: if the speaker/writer refers to an entity by means of an element at the top of the accessibility marking hierarchy, subsequent references to it can be done by using high accessibility markers. This is due to the fact that an informative, rigid, and unattenuated expression, that is, a low accessibility marker, temporarily heightens the salience of the referent.

The degree of accessibility of an entity or event also depends on the connection between the anaphoric element and its linguistic antecedent: the stronger it is, the higher the level of accessibility of the referent. First, this link is tightened if there is a short distance (in words, clauses, paragraphs, turns, etc.) between them. When the anaphor and its antecedent are close to each other, their connection is stronger and the accessibility of the antecedent increases. Second, the strength of the link is also contingent on the syntactic connection between the clauses in which the anaphor and the antecedent are located. If the clauses are independent syntactically speaking, then there is a looser connection between the two elements. On the other hand, if the clauses are connected by means of subordination or coordination, the link between the anaphor and its antecedent is stronger. Finally, if the two elements are located in different sentences, these can also be more or less tightly linked depending on the use of devices that increase pragmatic cohesion. For instance, if the transition from one sentence to the next is mediated by means of a connector that indicates a semantic relation between them,

such as *therefore*, *consequently*, or *so*, then the antecedent is more accessible to the anaphoric element.²⁴

Highly accessible entities or events, that is, those which are salient and tightly connected to the anaphoric elements that refer to them, are not difficult to process for the addressee because they can be easily retrieved from memory (Ariel 2001, 45–46). Therefore, they can be usually referred to by means of minimal forms – less informative, less rigid, and more attenuated – without incurring in extra processing costs, as long as the speaker correctly assesses the degree of accessibility of a referent in the addressee’s mind. Accessibility will figure prominently in the corpus study that is the focus of Chapter 5, in which this concept will be further elaborated on and will be operationalized for its use as a predictor of pronoun omission.

3.2.3 Further factors

Some other factors have also been shown to affect the omission of pronouns. Travis and Torres Cacoullos (2012) conducted a study of 1st person singular pronoun omission in subject position (*yo*) in spoken Colombian Spanish. They found that, besides accessibility, which they measure as the number of semantically compatible subject antecedents intervening between the (omitted/overt) pronoun and its correct referent (the larger the number of intervening subjects, the more ambiguous and, therefore, the less accessible the referent is), other variables influence the choice between an omitted and an overt pronoun. For one, subject omission in Colombian Spanish depends on the semantic class of the verb that follows: cognitive verbs (such as *saber* ‘know’ or *pensar* ‘think’) seem to favour overt expression,

²⁴ There is a further factor that may influence the connection between the anaphor and the antecedent, namely structural parallelism: if the anaphor and the antecedent are both subjects or objects, the link between them is tightened. However, the effect of this factor is less well understood than the rest (cf. Ariel 2001, 67).

while other verb classes do not. Additionally, overt pronouns are preferred in initial turn position but only within the class of cognitive verbs, not in other verb classes. Travis and Torres Cacoullos explain these findings in constructional terms, arguing that, within the more abstract SUBJECT PRONOUN + VERB schema, there is a specific *yo* + COGNITIVE VERB construction which has an interactional turn-taking function and exhibits a higher frequency of overt pronouns (cf. also Posio 2014 for constructional effects in Peninsular Spanish and European Portuguese).

A further factor identified in Travis and Torres Cacoullos (2012) is priming, or persistence in Szmrecsanyi's (2005) terms, which refers to the mechanical repetition of a recently used structural pattern. Two persistence effects were found in their study, as follows:

- the form of the previous coreferential subject: a coreferential overt *yo* pronoun favoured the overt expression of the subject, while an omitted pronoun was more likely if the previous coreferential subject was itself omitted.
- the form of the subject of the immediately preceding clause: an unexpressed pronoun selected *yo* omission more frequently; an overt pronoun favoured *yo* expression regardless of its person and number.

In addition, Travis and Torres Cacoullos (2012) found that the tense, aspect, and mood of the verb also affected the expression of 1st person singular pronouns: overt pronouns were more likely with imperfect, conditional, and subjunctive verbs and no effect was found with preterit and present forms. Previous claims in the literature (cf. Silva-Corvalán 1997, 2001) argued that imperfect, conditional, and subjunctive forms background the event expressed by the verb and thus favour overt subjects. On the other hand, omitted pronouns are more likely

with preterit verb forms, since they foreground the event. Finally, present tense forms are expected to play no role in the expression of subject pronouns. The results of Travis and Torres Cacoullos' study, however, do not support this hypothesis, since preterit forms have the same null effect as present tense ones. Their findings seem to reflect the influence of agreement on the expression of subjects: in Spanish, imperfect (e.g. *comía*, 'I/she/he/it was eating'), conditional (e.g. *comería*, 'I/she/he/it would eat'), and subjunctive (e.g. *comiese*, 'I/she/he/it ate', as in *I wish he ate*) 1st and 3rd person singular forms are identical, so agreement cannot unambiguously signal the morphological features of the subject in all person and number combinations and overt pronouns are more frequent.²⁵ Preterit and present tense paradigms, on the contrary, do not contain ambiguous forms, so they can successfully retrieve all the features of the subject in all person and number combinations and overt pronouns are not as common.

As pointed out, Travis and Torres Cacoullos' study (2012) is concerned only with 1st person singular pronouns, in isolation from other variables such as person and number. The influence of these, however, has been examined by Otheguy, Zentella and Livert (2007), who show that Spanish speakers in New York favour overt expression of 1st, 2nd, and 3rd person singular pronouns, while 1st and 3rd person plural forms are usually omitted (cf. Silveira 2007, 2008 for Brazilian Portuguese).²⁶ Therefore, it seems that it is also important to consider the

²⁵ The Spanish imperfect preterit tense is mainly used to present actions as they were taking place in the past, that is, without explicit reference neither to their beginning nor to their end (e.g., habitual actions, situations taking place at a past moment, circumstances surrounding a past event, and descriptions in the past; Real Academia Española 2009, 1688). The conditional is chiefly used to express non-factual information contingent on hypothetical situations (e.g., wishes, advice, past probability, and imaginary situations; Real Academia Española 2009, 1778). Finally, the subjunctive is employed in constructions that convey wishes, probability, or doubt, as well as other feelings (Real Academia Española 2009, chap. 24).

²⁶ 2nd person plural forms were not be studied by Otheguy, Zentella and Livert (2007) since they focused on Latin American varieties of Spanish in which a 3rd person plural form *ustedes* is more commonly used for 2nd person plural referents.

form of the pronoun that is being omitted. Claes (2017) argues that subject omission is influenced by the cognitive prominence of the entity being referred to, so that more prominent entities favour overt coding. Speakers tend to focus on themselves and their interlocutors rather than on other entities, meaning that 1st and 2nd person pronouns should be omitted less frequently than 3rd person pronouns. Claes finds this hypothesis to hold in Cuban Spanish, with subjects referring to speakers and hearers favouring overt expression.

Finally, Claes' (2017, 4) study also shows that subject omission in Cuban Spanish is subject to "statistical pre-emption," or "chunking" in the parlance of Haiman (1994, 8). This concept refers to the process whereby words or constructions that occur frequently together become stored in memory as a single component (Bybee 1998, 2001, 2002, 2006; Bybee and Scheibman 1999; Bybee and Thompson 1997; Krug 1998; Scheibman 2000). This new complex unit is also retrieved from memory and articulated in production as a whole, which sets off reductive processes within the boundaries of the unit that make it easier for speakers to process and produce. Krug (1998) has shown that the frequency of co-occurrence of a potentially contractible string of words is a factor affecting the incidence of contraction, so that the more frequently two words occur together, the more times they occur in a contracted form. Personal pronouns, which are high-frequency elements, are usually followed by verbs in a reduced form. For instance, the 1st person singular present tense form of the verb *be*, when preceded by the personal pronoun *I*, is commonly contracted onto the pronoun (*I'm*). Similarly, Bybee and Scheibman (1999) argue that the auxiliary *don't* occurs in a phonologically abbreviated form when it is part of a sequence of words that are frequently produced together, such as *I don't know*. The elements forming a complex unit are no longer treated as independent items when they occur within the new construction, that is, chunking implies the progressive loss of the internal constituent structure of the unit and the blurring of

the boundaries between its component parts. The individual items, however, still retain their own mental representations, which are stored in memory independently from the chunk. Bybee and Scheibman (1999, 301) postulate two cognitive mechanisms for the retrieval and production of complex units:

[...] we can postulate two cognitive mechanisms by which a phrase such as *I don't know* can be produced. In one case, the expression is a construction and accessed whole from storage, and it thus includes reductions and coarticulations that have accumulated in its representation; in the other case it is put together from two (*I don't* and *know*) or three (*I*, *don't*, and *know*) elements in which case the vowel of *don't* will not be reduced, though flapping can occur.

Claes (2017) finds that, when faced with the choice between omitted and overt subjects, speakers favour the variant that is more strongly associated with each particular verb form, thus relaying on their past linguistic experience. For instance, *yo creo* ('I think') is a "prefabricated unit" in Spanish according to Travis and Torres Cacoullos (2012, 739). Therefore, speakers do not commonly omit the subject pronoun *yo* when it is followed by *creo*, as these two elements are strongly associated with each other.

3.3 PRONOUN OMISSION IN ENGLISH

According to reference grammars (cf. Quirk et al. 1985; Biber et al. 1999; Huddleston and Pullum et al. 2002), pronoun omission in standard varieties of English occurs only in a restricted set of contexts. Besides being the default option in imperative sentences, subject pronouns are also omitted in the second conjunct of coordinate clauses, but only if they are coreferential with the subject of the first conjunct, as in (12).

- (12) **Sue_i** found the key and **Ø_i** unlocked the door. (Huddleston and Pullum et al. 2002, 1348)

Subject omission in coordination is a very common phenomenon in English and can be found in both speech and writing.

Another environment which favours the omission of subject pronouns is initial position in turns, sentences, and/or clauses (and sometimes non-clausal units), as in (13). This type of omission, however, is basically restricted to informal conversation or casual style and to main rather than subordinate clauses:

- (13) **Ø** Hope you are right. (Huddleston and Pullum et al. 2002, 1540)

Typically, the antecedents for these omitted subjects are recoverable from the situational context but, on some occasions, they are to be found in the previous discourse. In declarative sentences, 1st, 2nd, and 3rd person pronouns may be omitted in initial position, particularly at the start of a turn, as well as existential *there*, but the most frequently omitted subjects in these cases are *I*, dummy *it*, and *there*. When the subject is followed by an auxiliary verb, both can be deleted, especially when *I* and dummy *it* cooccur with a form of *be* (e.g. *Ø Glad you think so* or *Ø Strange how the ants come in when it's about to rain*; Huddleston and Pullum et al. 2002, 1541). In interrogative sentences, the subject can be omitted in initial position only if the preceding auxiliary is also unexpressed (e.g. *Ø Remember what you said the other day?*). The most common omitted element in these cases is *you*, although omission of other pronouns is also possible depending on the context. Subject omission in initial position is frequently

found in British English (in comparison with American English): for instance, according to Biber et al. (1999, 1105–6), in British English *Ø Depends* occurs more often than *It depends*.

Finally, subjects are deleted in a more widespread fashion in certain genres, such as diaries (Haegeman 1990) and product labels or labelese (Ruppenhofer and Michaelis 2010), and in reduced types of writing, namely headlines, titles, notes, postcards, etc.

As regards the omission of objects in English, it is contingent on particular genres and constructions: unexpressed objects occur in imperatives in instructional writing (14), in match reports (15), and in constructions containing quotative verbs that express judgement (16):

(14) Check **motor protection filter**_i every time you change the paper filter bag. Replace **Ø**_i by a new one if it is very dirty. (Ruppenhofer and Michaelis 2010, 162)

(15) Paramatti put **the ball**_i back into the box and Panadic headed **Ø**_i into the net. (Ruppenhofer and Michaelis 2010, 164)

(16) Nice work, **boys**_i, she praised **Ø**_i with a light smile. (Ruppenhofer and Michaelis 2010, 160)

Outside these environments, direct object pronouns are usually overt, although on some occasions they are omitted when their antecedents can be recovered from the linguistic or the situational context, as illustrated in (17):

(17) A: Show me **your essay**_i. B: I'll show **Ø**_i you later. (Quirk et al. 1985, 723)

Generally speaking, then, pronoun omission in standard varieties of English seems to be a restricted phenomenon that occurs only in certain registers, genres, and constructions.

However, this situation does not seem to apply to non-standard varieties. Among the 235 morphosyntactic features distinguished in *eWAVE*, the following five involve pronoun deletion: object pronoun drop (F42), as in (18), referential subject pronoun drop (F43), illustrated in (19), non-referential subject pronoun drop (F44), as in example (20), deletion of *it* in referential *it is*-constructions (F46), as in (21), and deletion of *it* in non-referential *it is*-constructions (F47), as in (22).

(18) I told him to get **some pears**_i but he forgot to bring **Ø**_i. (Example 3503; Indian South African English)

(19) When **I**_i come back from my work **Ø**_i just travel back to my home. (Example 902; Indian English)

(20) **Ø** Must be getting late. (Example 755; Sri Lankan English)

(21) My grandfather and grandmother speaks it and that's how I pick **it**_i up and **Ø**_i is still in me. (Example 21; Palmerston English)

(22) **Ø** Is raining. (Example 3291; Black South African English)

With the exception of F47, deletion of *it* in non-referential *it is*-constructions, which is attested in 28% of the varieties included in *eWAVE*, none of the features have an attestation rate lower than 30%: F42 is attested in 37% of the varieties, F43 in up to 51%, F44 in 36%, and F46 in 39%. These percentages show that referential subject omission (F43) is the most frequently attested feature, but the others are also quite common in non-standard varieties. Besides an attestation percentage, *eWAVE* also provides a measure of how pervasive each of

the features is in the varieties in which it occurs.²⁷ The pervasiveness ratings of the five pronoun omission features range from 56% (F46) to 65% (F42), meaning that they are neither extremely frequent nor extremely rare in the varieties in question.

3.3.1 Pronoun omission in the history of British English

The situation in Present-day English seems to be largely comparable to the situation in previous stages of the language; in other words, deleted pronouns have always had a marginal status in the grammar (cf. Walkden 2013; Rusten 2013, 2014, 2015; Walkden and Rusten 2017). In Old English (OE), referential omitted subjects occurred with an overall frequency of about 2% in prose texts and 11% in poetry. Therefore, it was already a very restricted phenomenon in OE, only occurring with a substantial frequency in poetry, probably for metrical considerations. There seems to have been regional variation already at this stage: according to Walkden (2013), Anglian (Northumbrian or Mercian) and Anglian-like OE texts, that is, those exhibiting Anglian features, displayed a greater percentage of referential deleted subjects (reaching 50% in some texts) than pure West-Saxon texts. In addition, a decrease in pronoun deletion is already noticeable over the period: early OE texts contain a larger proportion of deleted subjects than late OE. It appears, then, that in OE pronoun omission was

²⁷ eWAVE provides information on which features are attested in each of the 235 varieties included and how frequently they occur by means of the following ratings: ‘A’ (the feature occurs and is pervasive), ‘B’ (the feature occurs but is not pervasive), ‘C’ (the feature occurs but is rare), ‘D’ (the feature is not attested), ‘X’ (the feature is not applicable), and ‘?’ (no information is available). On the basis of these ratings, pervasiveness for each feature is calculated as the number of A values plus 0.6 times the number of B values plus 0.3 times the number of C values. The result is then divided by the number of varieties in which the feature is attested and multiplied by 100 to express it as a percentage. A pervasiveness of 30% means that the feature is rare in all the varieties in which it is attested, while a feature with a pervasiveness of 100% is very frequent in all the varieties in which it occurs. Attestation and pervasiveness will figure prominently in the cross-varietal study presented in Chapter 4, although pervasiveness will be calculated in a slightly different manner.

controlled mainly by extra-linguistic factors, namely genre, dialect, and subperiod, being common only in poetry, in Anglian texts, and in early OE.

As regards the linguistic constraints on the occurrence of omitted subjects in OE, it is noteworthy that they were possible in both main and subordinate clauses, although they were more common in the former than in the latter. Omission in the second of two coordinate clauses was especially frequent and, unlike in PDE, applied also in cases where the unexpressed subject of the second conjunct was *not* coreferential with the subject of the first conjunct (Walkden 2013; Rusten 2013, 2014, 2015). Furthermore, 3rd person pronouns were more commonly omitted than 1st or 2nd person ones. The antecedents of referential deleted pronouns were located in the immediately preceding clause more than 90% of the time, and they typically functioned as subjects rather than objects (or other grammatical roles). Interestingly, neither verbal morphology nor the discourse prominence of the antecedent, that is, whether the entity referred to was a topic in the discourse or not, played a decisive role in the identification of the referents of omitted subjects in OE: on many occasions, pronouns were unexpressed even though the verb with which they cooccurred did not signal the person and number of the subject unambiguously, and their antecedent was not a highly topical entity in the discourse. Finally, omitted objects and non-referential subjects were also attested in OE, although quantitative data on how frequently they occurred is, to the best of my knowledge, not available yet.

The decrease in the frequency of pronoun omission that can be observed already in OE continues in later stages. In Middle English (ME), referential omitted subjects were even more marginal than in OE, with a frequency of 0.6% in prose and a slightly higher one of 1.4% in poetry (Walkden and Rusten 2017). As regards regional and dialect provenance, it seems to have played no role at this stage. Along the chronological dimension, however, the date of the

text continues to have an impact: referential omitted pronouns occur more readily in earlier ME texts than in late ME. With respect to the linguistic factors conditioning pronoun omission, second conjunct clauses were the deletion environment par excellence, as in OE. Other than this, unexpressed subjects were also possible in both main and subordinate clauses, although more common in the former, again as in OE. Furthermore, the person and number of the pronominal form also influenced the occurrence of omitted subjects, with 2nd person singular and 3rd person pronouns being omitted more often than 2nd person plural and 1st person singular subjects. Finally, omission of non-referential subjects was also attested in ME.

By the early Modern English (eModE) period, referential deleted subjects had become a very marginal phenomenon, with an overall frequency of 0.5% (Rusten 2014), that is, lower than in OE and ME. Therefore, the evolution of pronoun omission throughout the history of English can be characterized as one of decline: it had a marginal status to begin with in the grammar of OE, and it further decreased in frequency in ME and eModE. Nowadays, as mentioned above, omission in standard varieties of English is restricted to certain registers and genres, and to a few specific constructions, although it seems to occur in a more widespread fashion in non-standard varieties.

3.3.2 Constraints in Present-Day English

Despite its low frequency and the restricted set of constructions in which it occurs, pronoun omission in English is a systematic phenomenon. The variation encountered between deleted and overt pronouns is not random but subject to both external and internal constraints. One of the most comprehensive studies of pronoun omission in a standard variety of English is Torres Cacoullos and Travis (2014). They focus on variable 1st person singular subject

expression in declarative main clauses in spoken American English (AmE), excluding 1st person plural, 2nd and 3rd person pronouns, non-finite clauses, interrogative sentences, subordinate clauses, and formulaic expressions such as discourse markers and quotative verbs (e.g., *I mean* or *I say*). They consider a variety of internal predictors, some of which were already included in Travis and Torres Cacoullos' (2012) study of 1st person singular subject omission in Colombian Spanish (cf. Section 3.2.3). In their 2014 study, they identify three significant constraints on subject expression in spoken AmE. First, an omitted subject is more likely in the second conjunct of a coordinate construction if it is coreferential with the subject of the first one, that is, the canonical environment for subject omission in English. Second, deletion is more common in the initial position of an intonation unit (IU), defined as a set of words uttered within the same intonation contour (Torres Cacoullos and Travis 2014, 25). Finally, persistence also has an effect: omission occurs more frequently when the previous coreferential subject is also unexpressed (cf. Gries 2005; Szmrecsanyi 2005, 2006 for further persistence effects in English).

Torres Cacoullos and Travis (2014) argue that the following three abstract schemas capture the phenomenon of 1st person singular subject omission in spoken AmE:

- (23) a. Coordination, [I_i VERB *and* $\emptyset_i$ VERB]: **I**_i went and **$\emptyset$** _i got a wet rag and wiped it off the car (Torres Cacoullos and Travis 2014, 26)
- b. IU initial position, [$\emptyset$ VERB ...]_{IU}: **$\emptyset$** Guess they gotta make money somehow (Torres Cacoullos and Travis 2014, 23)
- c. Persistence (both within and outside coordination contexts), [$\emptyset_i$ VERB (*and*) $\emptyset_i$ VERB]: last Sunday, **$\emptyset$** _i got there at eight, **$\emptyset$** _i left at ten, **$\emptyset$** _i dropped this person off at home (adapted from Torres Cacoullos and Travis 2014, 29)

Besides these three abstract schemas, omission is also found in various other more specific constructions. For instance, in coordination environments when both clauses occur within the same intonation unit ($[I_i \text{ VERB and } \emptyset_i \text{ VERB}]_{\text{IU}}$), which expresses a strong connection between the two events; this type, illustrated in (24), accounts for a substantial amount of Torres Cacoullos and Travis' (2014) data.

(24) I_i look at that and $\emptyset_i$ think (Torres Cacoullos and Travis 2014, 31)

The above type subsumes two more subtypes that are partly lexically filled: one with a quotative verb in the second conjunct, $[I_i \text{ VERB and } \emptyset_i \text{ QUOTATIVE VERB}]_{\text{IU}}$, as in (25a), and another with a verb of motion (typically *go*) in the first conjunct, $[I_i \text{ GO and } \emptyset_i \text{ VERB}]_{\text{IU}}$, as in (25b).

(25) a. I_i phoned her and $\emptyset_i$ said (Torres Cacoullos and Travis 2014, 31)

b. That's why I_i went out and $\emptyset_i$ got the coffee (Torres Cacoullos and Travis 2014, 31)

Other factors that were found to favour omission in Colombian Spanish in Travis and Torres Cacoullos (2012) do not affect 1st person singular subject expression in AmE. Accessibility, measured as the presence of intervening human subjects between the pronoun and the previous coreferential subject, is not a significant predictor of omission in Torres Cacoullos and Travis (2014). Likewise, turn position, polarity, semantic class of the verb, and

tense-aspect-mood do not influence the choice between an overt and a deleted subject pronoun.

Wagner (2018) examines subject pronoun omission in the variety of English spoken in the Canadian island of Newfoundland. The tokens analysed comprise only 1st person subject pronouns in declarative main clauses, excluding instances in coordination, enumerations, and lists, those followed by contracted forms of verbs (e.g., *I'm*, *I've*, and *I'll*), and examples in comment clauses (e.g., *I think*, and *I guess*). Wagner considers the influence of a set of extra- and intra-linguistic factors on the variation between 1st person omitted and overt subject pronouns, out of which the following six emerged as significant: priming or persistence effects, with omitted pronouns being favoured when immediately preceded by null subjects; verb phrase complexity, with complex verb phrases inhibiting the occurrence of omitted pronouns; turn position and length, with initial position and shorter turns favouring omission; verb semantic type, with perception verbs decreasing the probability of deleted pronouns; and, finally, switch reference, omission being favoured when the previous null subject has the same antecedent. The reader may recall that persistence and initial turn position were also significant predictors of pronoun omission in Torres Cacoullos and Travis' (2014) study on spoken AmE, which suggests that these two factors are important intra-linguistic constraints to be considered when explaining variation between omitted and overt pronouns across varieties of English. Extra-linguistic variables, on the other hand, such as the speakers' age, gender, or religious background, turned out not to have a significant impact on the probability of 1st person subject pronoun deletion in Wagner (2018).

Pronoun omission in English has also been approached from a generative perspective. An interesting proposal is that of Sato and Kim (2012), who focus on both subject and object deletion in Singapore English (SgE). On the basis of C.-T. J. Huang's (1984) theory of empty

categories (cf. Section 3.1), Sato and Kim argue that omitted subjects in SgE can refer back either to an empty element in topic position, or to the subject of the main clause if they occur in a subordinate clause. Deleted objects, on the other hand, can be coreferential only with an omitted topic. In Huang's theory, agreement, in those languages in which it is present, may function as an antecedent and thus help retrieve the person, number and gender features of an omitted subject. This, however, is not applicable to SgE: according to Sato and Kim, in this variety subject omission is blocked when the verb exhibits overt agreement with the subject, that is, with 3rd person singular present *-(e)s* and copula verbs. To account for this fact, they complement Huang's theory with that proposed by Speas (1994, 2006; cf. Section 3.1), who argue that weak agreement enters the derivation with unvalued morphological features and it thus requires the presence of an overt subject in the clause to identify them. Consequently, subject deletion is not allowed in weak agreement languages and, similarly, it is blocked in SgE in the presence of overt agreement.²⁸

A factor commonly put forward in order to explain the occurrence of omitted pronouns in high-contact varieties of English (cf. Chapter 2, Section 2.4) is transfer from the substrate language(s) (cf., among others, Bao 2001; Bayley and Santa Ana 2004; E. Schneider 2013). Along these lines, Schröter and Kortmann (2016) focus on subject pronoun omission in informal conversation in British English (BrE), Hong Kong English (HKE) and SgE, and argue that influence from the substrate languages generally accounts for the distribution found in the Asian varieties. They consider the expression of both referential and non-referential subjects with finite verbs in declarative and interrogative sentences, and in main and subordinate clauses, excluding imperatives, formulaic expressions, and cases of canonical

²⁸ Sato and Kim (2012), despite being an elegant proposal, is not an adequate description of pronoun omission in SgE since subject omission does occur with copula verbs (cf. Chapter 5, Section 5.2.2.2) and the proposed subject-object asymmetry does not hold in all cases (cf. Bao 2001).

omission in coordination (which are more frequent in BrE than in HKE or SgE). Overall, subject deletion is more common in SgE than in BrE, with HKE occupying an intermediate position. According to Schröter and Kortmann, the higher frequency of omission attested in the Asian varieties is due to the influence of their Chinese substrate, which allows covert expression of subjects and objects. Differences between SgE and HKE are explained by the status of English in Singapore and Hong Kong: SgE is an everyday language, thus implying a greater influence from the substrate than HKE. There are also important contrasts between the three varieties with respect to clause type: whereas omitted subjects in BrE occur almost exclusively in declarative main clauses, in SgE and HKE they can also be found in subordinate clauses, because the grammar of Chinese allows their occurrence in these environments. In interrogative sentences, there is a high frequency of deletion in SgE, especially in questions formed with the tag *or not*, as in (26), which is also a frequent structure in the southern Min Chinese dialects that are part of the Singaporean substrate:

- (26) Ø Remember or not? (ICE-SIN:S1A-023#91:1:B; quoted in Schröter and Kortmann 2016, 233)

In Mandarin and Cantonese, on the other hand, these constructions are not so common, thus explaining the absence of *or not*-questions in HKE. Non-referential subjects are also more frequently omitted in SgE and HKE than in BrE, since these are always unexpressed in Chinese: omission of dummy *it* is very common in HKE, and existential *there* is usually deleted in SgE due to the presence in this variety of the existential *got*-construction originated in the substrate languages:

(27) Here Ø got very many people. (Wee 2004, 1060)

Similarly, generic subjects are frequently omitted in the Asian varieties, and also in their Chinese substrates.

Another important issue to consider when examining pronoun omission in high-contact varieties of English is the simplifying effect of second language acquisition and use (cf. Chapter 2, Section 2.2.1). The literature is brimming with references to the occurrence of omitted pronouns in second-language speech in English and other languages (cf., among others, Dušková 1969; Butterworth and Hatch 1978; Felix 1980; Gundel and Tarone 1983; Huebner 1983; Dittmar 1984; Fakhri 1984; Gundel, Stenson, and Tarone 1984; Schumann 1984; Zobl 1984; Zyzik 2008). Williams (1988, 1989) compares the use of four referential devices, including pronoun copies (e.g., *That boy, he's smart*; Williams 1989, 162), definite noun phrases, and overt and omitted pronouns, in the speech of first-language users of AmE, second-language speakers of English with different first languages, and SgE speakers. She focuses on referential 3rd person subjects in declarative main and subordinate clauses and examines the effect of various accessibility measures on subject expression. Besides the fact that more explicit devices (i.e., pronoun copies and definite noun phrases) are used to refer to less accessible antecedents, there are also significant differences between the three groups of speakers. Overall, pronoun copies are more frequent in the production of second-language speakers, while omitted pronouns occur more commonly in second-language and SgE speech. First-language speakers use omitted subjects exclusively in coordination when the deleted pronoun refers back to the subject of the first conjunct. In second-language and SgE speakers' production, on the other hand, omission is also commonly found outside canonical coordination environments, for instance, when the antecedent occurs in the first conjunct but

is not the subject, or when the antecedent and the deleted pronoun are linked by subordination rather than coordination. Furthermore, the antecedents of omitted subjects in first-language speech are almost always located in the immediately preceding discourse, while in second-language and SgE speech they can be further removed from the unexpressed element. Finally, first-language and SgE speakers tend to use more explicit referential devices when other semantically compatible referents are found in the surrounding discourse so that there is ambiguity as to the correct antecedent of the anaphoric element. Second-language speakers, on the contrary, do not seem to take potential ambiguity into consideration when using an omitted subject.

The literature reviewed here shows that variation in the expression of pronominal elements in English is not arbitrary. Several factors, which will be further elaborated on and operationalized in Chapter 5, influence the choice between an omitted and an overt pronoun. This, in turn, means that, as a result of the availability of deleted pronouns, English has a grammar with more rules and, therefore, more complexity than it would if pronouns were categorically overtly expressed. Structures with omitted pronouns, however, are more economical than their overt counterparts, because they contain one fewer form that must be articulated and parsed. These issues related to the complexity of pronoun omission are the focus of the next section.

3.4 THE COMPLEXITY AND EFFICIENCY OF PRONOUN OMISSION

Any account of the complexity of pronoun omission needs to consider the following facts. First, as a result of the omission of subject or object pronouns, individual structures become formally simpler, that is, they contain fewer forms (and, possibly, also a simpler syntactic

representation if one does not commit oneself to the existence of empty categories). For instance, by comparison with example (28), (29) contains one more word:

(28) **Ø** Went to the cinema.

(29) **I** went to the cinema.

Second, the task of decoding the message is arguably more difficult for the addressee in the less explicit structure in (28) than in (29), since (28) requires access to the linguistic or situational context for the correct identification of the omitted subject.²⁹ Third, the availability of omitted pronouns in a grammar implies a larger set of referential expressions, since there is an extra variant in the set, namely pronoun deletion, in comparison with a language in which pronouns are invariably overtly expressed. And fourth, the existence of an extra referential device possibly entails a larger set of rules or constraints regulating its use. Sections 3.2 and 3.3 above showed that the occurrence of omitted pronouns is not random but systematic and, in fact, several factors seem to play a role in the choice between a deleted and an overt pronoun, which in turn involves a longer and more complex grammar.

What do these four facts mean for the complexity of pronoun omission? The reader may recall that, in Chapter 2, Section 2.3.5, some complexity distinctions were mentioned, namely global versus local, absolute versus relative, system versus structural, and overt versus hidden complexity. The last two distinctions will be briefly considered again here to properly define

²⁹ Nonetheless, it is debatable how much more transparent (29) is in comparison with (28), since overt pronouns, despite having unambiguous person and number features, also depend on the linguistic or situational context for the correct assignment of their antecedents.

and contextualize the metrics used in the rest of this section in order to assess the complexity of pronoun omission.³⁰

When studying the complexity/simplicity of a particular grammatical feature, it is important to take into account both systemic and structural issues, as focusing exclusively on one of these perspectives may render an incorrect, or at least incomplete, assessment of its contribution to the complexity of a language or dialect. Recall that whereas system complexity metrics concentrate on the complexity of paradigms and rules mapping meanings to forms, structural complexity metrics measure the outputs of those rules, that is, the actual linguistic structures. Additionally, within system complexity, a further division can be made between *tectogrammatics* and *phenogrammatics*: the former refers to the different structural choices in a grammar, while the latter deals with the rules that constrain the use of those choices (in both cases the larger the number, the more complex the system). It is possible for a grammatical feature to result in simpler and efficient linguistic structures while simultaneously complexifying the system by means of additional structural choices and rules. In fact, as discussed below, this is what happens in the case of pronoun omission.

Similarly, the analysis of a grammatical feature should be informed by both overt and hidden complexity considerations: whereas overt complexity refers to the explicit morphosyntactic mechanisms used to signal grammatical categories and rules, hidden complexity involves the covert processes of inference that are necessary for addressees to successfully retrieve implicit information. It is not only possible but likely that a grammatical feature that decreases the complexity of a language or dialect in overt terms results in an

³⁰ In the present monograph, complexity will be approached exclusively from a local and relative perspective for the reasons explained in Chapter 2, Section 2.3.5. Therefore, neither the distinction between global and local complexity nor that between absolute and relative approaches will be discussed here again.

increase in hidden complexity, given that at least some unexpressed information is bound to give rise to inferential processes on the part of addressees.

With these issues in mind, we are now in a position to evaluate the complexity of pronoun omission. From a structural perspective, pronoun omission results in constructions that are easier to process. This is because it minimizes their formal complexity (cf. Hawkins' MiF principle in Chapter 2, Section 2.3.3) and allows for an earlier access to their ultimate syntactic and semantic representation (cf. Hawkins' MaOP principle).

First, formal complexity is minimized because there is an extra form that must be articulated and processed when pronouns are overt, so pronoun omission makes structures more economical for speakers. This, in turn, does not necessarily entail that hearers need to make an increased effort to correctly assign a referent to the omitted element as long as they have access to sufficiently specified contextual information, broadly construed.³¹ High entity or event accessibility in the discourse helps identify the referents of deleted pronouns, as well as the presence of an antecedent in the linguistic context with which the omitted element can be easily coindexed (cf. Section 3.2.2). Moreover, reference identification can also be facilitated by means of structural factors such as agreement between the verb and the deleted pronoun (cf. Section 3.2.1), or constructional parallelism between the clauses containing the antecedent and the omitted anaphoric expression (as in cases of coordination when the omitted subject is located in the second conjunct and refers back to the subject of the first one).

Second, pronoun omission allows the hearer to access the ultimate syntactic and semantic representation of the sentence earlier than when pronouns are overt. Hawkins' MaOP

³¹ This is not an issue in cases of omission of non-referential pronouns, which only play a structural rather than a semantic role in the clause.

principle postulates that users prefer to assign the larger number of properties possible to each linguistic form as each of these forms is encountered in online processing, thus increasing the amount of syntactic and semantic structure that can be correctly parsed at each step of the process. Tables 1 and 2 show the number of properties that can be identified as each word is encountered in the online processing of examples (28) and (29) above respectively (cf. Hawkins 2004, 55–58):³²

Table 1. Analysis of example (28)

	$\emptyset$	<i>Went</i>	<i>to the cinema</i>
Categories	-	V ₁	...
Phrases	-	VP ₁ , S ₁	...
Attachments	-	vp1[V ₁], s1[VP ₁]	...
Relations	-	$\emptyset = \text{SU of } V_1$	...
Prop. ratio	-	6/6 = 100%	...

Table 2. Analysis of example (29)

	<i>I</i>	<i>went</i>	<i>to the cinema</i>
Categories	Pro ₁	V ₁	...
Phrases	NP ₁ , S ₁	VP ₁	...
Attachments	NP1[Pro ₁], s1[NP ₁]	vp1[V ₁], s1[VP ₁]	...
Relations	-	NP ₁ = SU of V ₁	...
Prop. ratio	5/10 = 50%	10/10 = 100%	...

³² Tables 1 and 2 present the analysis of the parts of examples (28) and (29) that are relevant to the present discussion, that is, the subject and the verb.

The leftmost columns of Tables 1 and 2 contain the different types of properties: the grammatical category of each word, the phrases that can be construed, the attachments of words to phrases that can be established, and the syntactic/semantic relations between words that can be set up at each step. The last row shows the number of properties identified as each item is processed, the total number of properties in the relevant part of the sentence, and the ratio of correctly identified properties at each step, expressed as a percentage. The first word the hearer encounters in example (28) is the verb *went* because the subject pronoun is omitted. Six out of a total number of six properties in the sentence can be processed at this stage, that is, a ratio of 100%: *went* is a verb (V_1), so a verb phrase can be construed (VP_1) and a clause over it (S_1); V_1 can be attached to VP_1 ($VP_1[V_1]$) and this, in turn, to S_1 ($s_1[VP_1]$); finally, the subject of V_1 is an omitted pronoun ($\emptyset = SU$ of V_1), which can be coindexed with an antecedent in the linguistic or situational context. In example (29), on the other hand, the first word is *I*. Five features can be identified as this word is processed: *I* is a pronoun (Pro_1), so a noun phrase can be construed (NP_1) and a clause over it (S_1); Pro_1 can be attached to NP_1 ($NP_1[Pro_1]$) and this, in turn, to S_1 ($s_1[NP_1]$). The second word is *went*, and another five properties can be parsed at this stage: *went* is a verb (V_1), so a verb phrase can be construed (VP_1); V_1 can be attached to VP_1 ($VP_1[V_1]$) and this, in turn, to S_1 ($s_1[VP_1]$); the subject of V_1 is NP_1 ($NP_1 = SU$ of V_1), which is realized by an overt pronoun that can be coindexed with an antecedent in the linguistic or situational context. Out of a total number of ten properties in the relevant section of example (29), only five can be parsed when *I* is encountered, a ratio of 50%, meaning that the hearer has to wait until *went* to have access to the ultimate syntactic and semantic representation of this part of the sentence. In example (28), this can be achieved

in the very first word, the verb *went*. Therefore, pronoun omission also results in easier-to-process structures according to MaOP.³³

Finally, there are no changes between omitted and overt pronouns as regards Hawkins' MiD principle, since the same coindexation relation is established between both types of referential devices and their respective antecedents, and the domain required in order to identify this dependency is neither increased nor decreased as a result of pronoun omission.

In summary, structures with omitted pronouns are economical and provide the hearer with early access to their ultimate syntactic and semantic representation, that is, they are easier to process than their overt counterparts. Crucially, this decrease in overt complexity is not necessarily, or at least not in all cases, at the expense of hidden complexity. It is true that, as discussed above, the lesser explicitness of structures like that in (28), in comparison with (29), is likely to give rise to processes of pragmatic inference on the part of addressees in order to assign a referent to the omitted element. However, contextual information, broadly construed, can be resorted to so as to make the search for an antecedent easier. As mentioned in Section 3.2.2, accessible entities or events in discourse are easy for addressees to retrieve so, in cases in which the referent of an omitted pronoun is highly accessible, the increase in hidden complexity is partially compensated for. Pronoun omission also allows for a more efficient transmission of information in Hawkins' terms (cf. Chapter 2, Section 2.3.3): communication is efficient when the speaker is able to transmit a message to the addressee with minimum processing effort and within the shortest possible time frame. Structures with omitted pronouns are easier to process and they contain fewer forms that must be articulated than those with overt pronouns, meaning that speakers can transmit information more efficiently (faster and with less effort).

³³ Even though (28) and (29) are instances of subject pronouns, the same holds in cases of omitted objects.

Further evidence that pronoun omission results in simpler, easier-to-process structures comes from its frequent occurrence in second-language production (in English and other languages) mentioned in the previous section, since, as discussed in Chapter 2, Section 2, simplification is a consequence of language contact and, particularly, short-term adult acquisition. Additionally, pronoun omission should also be favoured by learners of a second language according to Filipović and Hawkins' (2013) principles: deleted pronouns, especially if they are allowed in the learner's first language, are beneficial for MiL and MiP, because simpler structures minimize learning and processing effort, and for MaC, since they also make communication more efficient.

From a systemic perspective, on the other hand, pronoun deletion increases the complexity of the grammar. Tectogrammatical complexity increases due to the existence of an extra referential device, namely omitted pronouns. Similarly, the phenogrammatical complexity of the system grows because of the extra rules or constraints needed to account for the variation between omitted and overt pronouns (and other referential expressions). Larger sets of referential expressions and more rules to regulate their use complexify the system and, consequently, increase the effort required on the part of speakers to acquire it. This extra systemic complexity should also be dispreferred by second-language speakers according to Filipović and Hawkins' MiL principle, since the omission of pronouns arguably results in more learning effort on their part.

Pronoun deletion, then, generates economical structures which also allow for a more efficient transmission of information, though it seems to be at the expense of a more complex system with a larger set of referential expressions and, possibly, more rules that regulate the variation between omitted and overt pronouns. This increase in system complexity, however, might be mitigated if the set of factors that constraint the use of omitted pronouns is minimal,

so as not to incur in severe extra learning difficulties. If this is true, speakers, and especially second-language ones, might still favour the omission of pronouns whenever possible due to its potential for form minimization, which in turn could be reflected in a higher attestation of this feature in high-contact varieties of English. As argued by Williams (1988, 341):

If the referents are recoverable, it is possible that repeated explicit reference is considered redundant by the speaker. If the speaker's production is already constrained by limited competence, this kind of omission of redundant elements is an attractive short cut. In an effort to keep production as economical as possible, constituents such as pronouns may simply be omitted with little loss of meaning.

Chapters 4 and 5 will present two studies that put these hypotheses to the test. Chapter 4 will focus on whether pronoun omission is attested more frequently in high-contact varieties of English than in low-contact ones by means of a cross-varietal survey of pronoun deletion in the varieties of English included in *eWAVE*. Chapter 5, on the other hand, will present a corpus-based study which examines the factors that explain the variation between omitted and overt pronouns in two high-contact varieties, namely SgE and Indian English (IndE), and in BrE, a low-contact one.

4 A CROSS-VARIETAL STUDY OF PRONOUN OMISSION IN ENGLISH

Pronoun omission, despite not being a completely innovative grammatical feature in the sense that it is also found in standard varieties of English, manifests a clearly different distribution in non-standard varieties, where it seems to occur more frequently and in a wider range of constructions. However, two questions that remain to be properly answered are why pronoun omission is attested in some varieties but not in others, and why it is more pervasive in some varieties than in others. For instance, deletion of referential subject pronouns is present in the grammar of 39 out of the 76 varieties included in *eWAVE* and, out of the varieties in which it occurs, it is highly frequent in eight, neither frequent nor rare in eighteen, and rather rare in thirteen. The presence of morphosyntactic innovations in varieties of English has commonly been explained as a result of one (or both) of two factors; these are simplification processes due to second-language acquisition and use, and influence from the substrate language(s) that English is/was in contact with. Substrate influence may in turn manifest itself as (i) the addition of completely novel features to the grammar of a variety following the template of the substrate language(s), or (ii) in the form of statistical preferences, that is, the under/over-production of certain grammatical patterns in the second language under the influence of the first language (cf. Brunner 2017, 35; Huber 2012). As mentioned in Chapter 3, Section 3.3.2, pronoun omission has traditionally been considered a clear example of transfer or

reinforcement from substrate languages, but it is also characteristic of the speech of second-language users of various languages and with different first-language backgrounds. This points to the possibility of pronoun omission also being the result of simplification due to second-language acquisition and use since it generates economical and efficient structures with fewer overt elements.

One caveat against arguing for a monofactorial account of any innovative feature is the difficulty of teasing apart the effects of the two factors because, in many cases, the features that are transferred can also be categorized as simple. Consider the use of preverbal negator *no* (F160 in *eWAVE*), a characteristic feature of pidgins and creoles, as a case in point:

(30) Me **no** know. (Example 2715; Trinidadian Creole)

Many of the substrate languages of English-based pidgins and creoles have an invariant preverbal negator marker (A. Schneider 2012, 888), which suggests that its presence in these varieties is due to substrate influence. However, Szmrecsanyi and Kortmann (2009a, 70) claim that preverbal negator *no* eases second-language acquisition and use, as evidenced by its frequency in learner speech. In this case and many others, simplification and substrate influence are highly intertwined processes, which means that advocating for an explanation based on just one factor runs the risk of telling only part of the story.

Against this backdrop, and acknowledging the correlation between simplification and substrate effects (cf. Biewer 2015, 83; Brunner 2017, 24), the present chapter aims at examining the explanatory power of each of the two factors with respect to the attestation and pervasiveness of pronoun omission features in varieties of English. Rather than favouring an account based on either simplification or substrate effects, the study presented below is an

attempt to ascertain which one is a better predictor of pronoun omission, while at the same time assuming that both are necessary to reach a fuller understanding of this feature. This will be done by conducting a cross-varietal survey of the distribution of the five pronoun omission features in *eWAVE* introduced in (18)-(22) above and repeated in (31)-(35) for convenience:

(31) Object pronoun drop (F42): I told him to get **some pears_i** but he forgot to bring **Ø_i**.

(Example 3503; Indian South African English)

(32) Referential subject pronoun drop (F43): When **I_i** come back from my work **Ø_i** just travel back to my home. (Example 902; Indian English)

(33) Non-referential subject pronoun drop (F44): **Ø** Must be getting late. (Example 755; Sri Lankan English)

(34) Deletion of *it* in referential *it is*-constructions (F46): My grandfather and grandmother speaks it and that's how I pick **it_i** up and **Ø_i** is still in me. (Example 21; Palmerston English)

(35) Deletion of *it* in non-referential *it is*-constructions (F47): **Ø** Is raining. (Example 3291; Black South African English)

The degree of attestation and pervasiveness of the five features will be explored across regions and variety types so as to discover which geographical areas and socio-historical conditions favour their occurrence. These findings will be then related to the two explanatory factors: the effects of variety type will be accounted for on the basis of simplification processes due to second-language acquisition and use, while geographical patterns will be connected to the influence of the substrate languages spoken in each region. Variety type refers to “the socio-historical conditions in which a given variety emerged” (Kortmann 2013,

166) including, among others, the degree of contact with other languages and/or dialects of English, their prestige in the speech community, and, more importantly for the purposes of the present work, their status as first or second languages, with the latter typically exhibiting grammatical simplification in comparison to the former. On the other hand, regional patterns with respect to the distribution of features in varieties of English are commonly interpreted as a result of transfer from the substrate languages that English entered into contact with (cf. Kortmann 2013): the fact that a feature is present in some regions and not in others is considered to be a consequence of its occurrence or absence in the relevant substrate languages.

The rest of the chapter is structured as follows. Section 4.1 reviews the evidence in favour of accounting for pronoun omission features in English on the basis of simplification processes and substrate influence. Section 4.2 introduces the data and methodology employed in the cross-varietal study, followed by the results in Section 4.3. Finally, in Section 4.4, the findings are related to the theoretical background of the monograph.

4.1 SIMPLIFICATION AND SUBSTRATE EFFECTS

As shown in Chapter 3, Section 3.3.2, the attestation of pronoun omission features in contact varieties of English has traditionally been explained as a consequence of transfer or reinforcement from substrate languages. However, in the second-language acquisition literature, omitted pronouns have been found to occur frequently in the speech of second-language learners, irrespective of their first-languages. Moreover, in Chapter 3, Section 3.4, it was argued that pronoun omission simplifies structures by making them more economical and, therefore, it should be favoured by second-language learners (as long as there is not a substantial increase in the complexity of the system). We thus encounter two possibly

contradictory scenarios: (i) pronoun omission in contact varieties of English is (mainly) a result of substrate influence, or (ii) it is (mainly) a result of simplification due to second-language acquisition and use. In addition, we can also try to reconcile these two positions by assuming that both factors may in principle play a role, the effects of which are in fact very difficult to tease apart (cf. the discussion of example (30) above).

If we consider the varieties' overall morphosyntactic profiles, region is of secondary importance as an explanatory factor in comparison with variety type. More structurally coherent clusters arise by grouping varieties according to their type than if they are grouped according to the world region in which they are spoken, that is, type provides a much more robust account than geography of the co-presence or co-absence of grammatical features across varieties (Kortmann 2013, 171). Region, and thus substrate influence, has a much stronger effect if we focus on individual features rather than on overall morphosyntactic profiles. Kortmann (2013) envisages three possible scenarios in which region may be more important as an explanatory factor. First, it may explain the degree of pervasiveness of a feature or set of features in a specific group of varieties. Second, the degree to which a certain grammatical tendency is observed in individual varieties may also be related to the substrate languages spoken in the geographical area in which they are located: for example, Asian varieties show a tendency to delete grammatical elements due to the preference for omission characteristic of their substrate languages, while African varieties favour their preservation or even the insertion of new ones because of the preserving nature of sub-Saharan African languages (cf. Mesthrie and Bhatt 2008, 90–91). Third, substrate influence may also account for the use of specific constructions or lexical items for the encoding of a grammatical function in a particular variety.

However strong the effect of region may be in those three scenarios, it is important to note that substrate influence is not explanatory on its own. As argued by Mufwene (2008, 150), the mere presence of a grammatical feature in the substrate language(s) does not inevitably entail its selection over others: many features are never transferred, and others may occur at a certain point but never find their way into the communal norm. Reference must necessarily be made to other factors that account for the selection of some features over others in particular contact situations, such as complexity (simple features are favoured; complex ones are disfavoured), frequency (frequent features are preferred over infrequent ones), saliency (salient features are more easily transferred than non-salient ones), markedness (unmarked features are selected over marked ones), or congruence between the substrate and the superstrate (features that are compatible with the grammar of the superstrate language are more readily transferred; cf. also Bao 2010).

Schröter and Kortmann (2016) argue for a substratist explanation for the occurrence of pronoun omission in varieties of English. They identify several features that are closely connected with particular geographical regions by comparing their attestation rates in a particular area with their rates in the rest of the Anglophone world. In the case of Asia, they find sixteen features that are very frequently attested in Asian varieties but are not so common in other regions, with seven of these features describing instances of omission of grammatical elements and four out of these seven being pronoun omission features. Schröter and Kortmann connect these findings to the substrate languages of the Asian region, which are characterized by a tendency towards the omission of grammatical elements (Mesthrie and Bhatt 2008, 90–91; cf. above). Further evidence of the relation between substrate influence and pronoun omission in contact languages comes from the *Atlas of Pidgin and Creole Language Structures Online* (APiCS; Michaelis et al. 2013) and the *World Atlas of Language*

Structures Online (WALS; Dryer and Haspelmath 2013). Schröter and Kortmann find that there is a clear geographical distribution of subject pronoun omission (F62 in *APiCS*; F101A in *WALS*). Omitted subject pronouns are more frequently attested in pidgins and creoles located in the Indo-Pacific and Asian regions, and this distribution essentially overlaps with the one found in the languages included in *WALS*, with a higher attestation rate in Australo-Pacific and Asian languages. On the basis of these distributions and the fact that pronoun omission is not particularly frequent in pidgins and creoles (only eighteen out of 76 pidgins/creoles in *APiCS* exhibit optional subject pronouns), they argue that region, and thus influence from the substrate languages, is a better predictor of the occurrence of deleted pronouns in English.

However, in light of the evidence discussed in this section, we can hypothesize that both region and variety type are necessary in order to reach a complete and coherent explanation of the (non-)attestation and pervasiveness of pronoun omission in varieties of English. First, despite exhibiting a clear geographical distribution, the mere presence of a feature in the relevant substrate languages does not suffice as an account of its attestation and pervasiveness in contact varieties, because this does not explain its selection over other features available to speakers in particular contact situations; other factors, such as complexity, must be resorted to. Second, substrate influence and simplification due to second-language acquisition/use are highly intertwined factors and, in many cases, it is extremely difficult or even impossible to distinguish their independent effects. Third, simplification and substrate influence may apply at different levels, with the former accounting for the varieties' overall morphosyntactic profiles and the latter accounting for the degree of pervasiveness of individual features, their presence or absence in specific varieties, and the use of different constructions and lexical items to encode the same grammatical function across varieties. In what follows, a cross-

varietal study is presented which aims at assessing the explanatory power of region (functioning as a proxy for substrate influence) and variety type (functioning as a proxy for simplification effects) with respect to the attestation and pervasiveness of pronoun omission features in varieties of English.

4.2 DATA AND METHODOLOGY

The main goals of the present chapter are (i) to explore the distribution of pronoun deletion features across world regions and variety types, and (ii) to measure the importance of simplification and substrate influence as explanatory factors for the occurrence of omitted pronouns in varieties of English. In other words, the aim is to answer the following research questions: in which world regions and variety types is pronoun omission more pervasive and more frequently attested? Which variable, region or variety type, is a better predictor of pronoun deletion features in the Anglophone world? To this purpose, a cross-varietal study was conducted using *eWAVE* as a source of data, which contains information on the attestation and pervasiveness of 235 morphosyntactic features of spontaneous speech in 76 varieties of English. For each feature, *eWAVE* informs of its attestation and pervasiveness in each of the varieties by means of six different values:

- A: the feature occurs pervasively or even obligatorily in the relevant variety.
- B: the feature is neither highly frequent nor rare.
- C: the feature occurs but is extremely rare.
- D: the feature is absent from the grammar of the relevant variety.
- X: the feature is not applicable.

- ?: there is no information as to the attestation and pervasiveness of the feature in the relevant variety.

On the basis of the ratings of the five pronoun omission features exemplified in (31)-(35) above (i.e., F42, F43, F44, F46, and F47 in *eWAVE*), two sets of indexes were computed. The first one measures the attestation of pronoun omission features in a particular variety: they are attested if they are given an A, B, or C value and not attested otherwise. The second one quantifies the pervasiveness of pronoun omission in those varieties in which it is attested, that is, how frequently omitted pronouns occur. Following the established practice in *eWAVE*, this second set is calculated by transforming A, B, and C values into numbers, namely 1, 0.6, and 0.3 respectively. As illustrative examples, consider Tables 3 and 4, which contain the attestation and pervasiveness indexes for White Zimbabwean (WhZimE) and White South African English (WSAfE) respectively:

Table 3. Attestation and pervasiveness in White Zimbabwean English

WhZimE	Attestation	Pervasiveness
Object omission	Not attested (D)	-
Ref. subject omission	Not attested (?)	-
Non-ref. subject omission	Not attested (D)	-
Deletion of ref. <i>it</i>	Not attested (D)	-
Deletion of non-ref. <i>it</i>	Not attested (D)	-
Pronoun omission	Not attested	-

Table 4. Attestation and pervasiveness in White South African English

WhSAfE	Attestation	Pervasiveness
Object omission	Attested (B)	0.6
Ref. subject omission	Attested (B)	0.6
Non-ref. subject omission	Not attested (?)	-
Deletion of ref. <i>it</i>	Not attested (D)	-
Deletion of non-ref. <i>it</i>	Not attested (D)	-
Pronoun omission	Attested	1.2

In WhZimE none of the features has an A, B, or C value, which means that they are not attested and we do not calculate their pervasiveness. In WhSAfE, on the other hand, object and referential subject omission are attested, since they are both given B values, and thus we can calculate their pervasiveness by transforming the values into numbers, namely 0.6. The final row in Tables 3 and 4 contains two global indexes that attempt to provide an overall measure of the attestation and pervasiveness of pronoun omission in the varieties. Pronoun omission attestation is computed as follows: it is attested if at least one of the five features receive an A, B, or C rating in the relevant variety and is not attested otherwise. Pronoun omission pervasiveness is calculated by summing the individual pervasiveness scores of the attested features. In WhZimE (Table 3), pronoun omission is not attested, since none of the five features is, so its global pervasiveness is not calculated. In WhSAfE (Table 4), pronoun omission is attested (two features are given B values) and it has a pervasiveness of 1.2, which is the sum of the individual pervasiveness scores of the two attested features. This further step of bundling all pronoun omission features together in two global indexes is carried out with the aim of making the most of the available *eWAVE* data: not many varieties attest some of the features (for instance, the deletion of non-referential *it* in *it is*-construction, F47, is attested

in only 28% of the varieties in *eWAVE*), so statistically significant results are hard to obtain in some cases. However, the five pronoun omission features are very heterogeneous, so the findings from the two global indexed should be taken with care.

To measure the effects of variety type and region with respect to the attestation and pervasiveness of pronoun omission features, the varieties were classified according to their type and the geographical area in which they are spoken. To this purpose, the classification provided in *eWAVE* was followed, which distinguishes between eight regions (i.e., British Isles, America, Africa, Asia, Australia, Pacific, Caribbean, and South Atlantic) and five variety types (i.e., traditional low-contact L1s, high-contact L1s, indigenized L2s, pidgins, and creoles).³⁴ The two L1 categories differ in the amount of contact with other languages and/or dialects which they underwent and in their internal consistency. Traditional L1s comprise regional non-standard varieties that have not entered into contact with other dialects or languages in the last 400 years. The high-contact L1 category, on the other hand, is the most heterogeneous of the types, as it includes varieties which are the result of dialect- and/or language-contact, namely transplanted L1s, standard L1s, and language-shift varieties.³⁵ The other three categories include non-native varieties of English (i.e., indigenized L2s and pidgins) and native varieties that were formed in situations where a group of non-English speakers had to acquire the language with restricted access to native speakers (i.e., creoles).

³⁴ The South Atlantic region is included here for the sake of completeness, but it does not really represent a coherent geographical area due to the large distance between the three islands in which the South Atlantic varieties are spoken (Kortmann 2013, 165–66). Therefore, the distribution of pronoun omission features in South Atlantic varieties will only be commented on in passing.

³⁵ Transplanted L1s are varieties that developed in former settlement colonies over the last 400 years as a result of dialect- and/or language-contact. Shift L1s are those which were mostly used as L2s in the past but that have been adopted as L1s by the majority of their speech community.

In line with Kortmann (2013, 186), potential geographical patterns will be considered here to derive from the influence of the substrate languages spoken in each of the regions identified. However, for variety type to reflect simplification effects due to second-language acquisition and use, the high-contact L1 category had to be further refined, since it includes both L1 varieties that were or are still affected by contact with other languages and varieties that are primarily the result of dialect contact. Therefore, high-contact L1s were classified into dialect-contact and shift L1s, resulting in a total of six variety types which clearly separate varieties which are (or were) affected by the process of second-language acquisition/use from those that are not. Table 5 provides the classification of the 76 varieties included in *eWAVE* according to their region and type.³⁶

In order to disentangle the effects of simplification and substrate influence and measure their relative impact on the attestation and pervasiveness of pronoun omission features in varieties of English, the data was fed into regression analyses (cf. Cysouw 2013). Regression modelling is a statistical technique that aims at explaining the relationship between a dependent variable and one or more independent variables, also called predictors. Linear regression is used to model the relationship between a numeric dependent variable and one or more categorical and/or numeric predictors (Levshina 2015, 139), while logistic regression estimates the likelihood of the levels of a categorical dependent variable on the basis of a series of categorical and/or numeric predictors (Baayen 2008, 195; Levshina 2015, 253). The reason why regression modelling is used in the present chapter to measure the explanatory power of variety type (as a proxy for simplification) and region (as a proxy for substrate

³⁶ However, variety type and region are only crude proxies of simplification due to second language acquisition/use and substrate effects, respectively. Variety type involves much more than the acquisition process (cf. Kortmann 2013, 166), and grouping the varieties into macro-regions cannot possibly capture the whole linguistic diversity characteristic of the countries in which they are spoken (cf. Fuchs 2016, 249).

influence) is that it enables us to estimate the effect of each predictor on the dependent variable while controlling for the impact of the other predictors in the model.

Table 5. Classification of varieties according to region and variety type³⁷

Type Region	Traditional L1s	Dialect- contact L1s	Shift L1s	Indigenized L2s	Pidgins	Creoles
British Isles	OSE, North, SW, SE, EA, ScE		IrE, ManxE, WelE, ChIsE, MaltE			BrC
America	NfldE, AppE, OzE, SEAmE	CollAmE	UAAVE, RAAVE, EAAVE	ChcE		Gullah
Africa		WhZimE, WhSAfE	LibSE	GhE, NigE, CamE, KenE, TznE, UgE, BlSAfE, InSAfE, CFE	VLibE, GhP, NigP, CamP	Krio
Asia			CollSgE, PhE	IndE, PakE, SLkE, HKE, MalE	ButLE	
Australia		AusE, AusVE	AborE			RRC, TorSC
Pacific		NZE		CollFijiE, FijiE	Norfk, TP	PalmE, Bisl, HawC

³⁷ The varieties are abbreviated following Kortmann (2013).

Caribbean			BahE	JamE		BahC, Bajan, JamC, SanAC, BelC, GuyC, EMarC, Saram, Sranan, TrinC, VinC
South Atlantic		FlkE, TdCE	StHE			

Logistic regression analyses were computed by means of the `glm()` function to measure the influence of variety type and region on the (non-)attestation of pronoun omission features, while linear regressions, with the pervasiveness indexes as dependent variables, were fitted using the `lm()` function.³⁸ The importance of variety type and region as predictors of the attestation and pervasiveness of pronoun omission features was evaluated by means of the `drop1()` function, which eliminates predictors from the model one at a time and assesses the differences with the model containing all predictors. The output of this function includes a significance value for each predictor which determines whether removing a given predictor results in a significantly worse model. In addition, `drop1()` gives the Akaike Information Criterion (AIC) computed after eliminating predictors from the model. AIC is a measure of goodness-of-fit useful for model comparison – the smaller the value, the better the fit – which penalises models that include many explanatory variables (Levshina 2015, 149). Non-significant predictors were removed from the formula and models containing only significant predictors were computed to obtain the coefficients, that is, the models' predictions for the dependent variable (here, either attestation or pervasiveness).

³⁸ All the statistical calculations carried out in this monograph were performed using *R* (R Core Team 2017).

Finally, following Cysouw (2013), I inspected the residuals after regression with either variety type or region as predictors of pronoun omission attestation and pervasiveness. Residuals are calculated by subtracting the predicted values of the dependent variable computed by the model from the actual observed values: positive values indicate that the predictions of the model are too low, while negative values indicate that they are too high. The idea is to ascertain whether residual regional or variety-type patterns can be observed in the data after controlling for the influence of the other predictor, which could in turn be used to explain the presence of outliers, that is, varieties that exhibit a value that is too high or too low for their variety type or region.

4.3 RESULTS

Two different sets of findings are reported in this section, namely those related to the global indexes of pronoun omission attestation and pervasiveness (Section 4.3.1) and the results of the individual pronoun omission features (Section 4.3.2). As regards the overall degree of attestation of pronoun omission, it is more frequently attested in shift L1s, indigenized L2s, pidgins, and creoles than in traditional and dialect-contact L1s. Region, on the other hand, does not have a significant effect on the rate of attestation but it does influence the pervasiveness of pronoun omission in those varieties in which it is attested, since it occurs more pervasively in Asia and the Pacific. With respect to the attestation of each of the individual features, only region is significant, and it accounts for the distribution of four of them: object omission is more frequently attested in Asia and the Pacific, non-referential subject omission is commonly found in Asian, Pacific, and Australian varieties, deletion of referential *it* in *it is*-constructions is characteristic of the Caribbean, and, lastly, deletion of non-referential *it* in *it is*-constructions is frequently attested in Asia and the Pacific. Region

also affects the pervasiveness of two pronoun omission features, namely object omission, which is pervasive in Asia, Australia, and the Pacific, and deletion of non-referential *it*, which is characteristic of the Caribbean and, to a certain extent, Asia and the Pacific. Section 4.3.1 deals first with the global indexes of attestation and pervasiveness.

4.3.1 Global indexes: pronoun omission attestation and pervasiveness

As shown in Figure 1, two groups of variety types emerge in relation to the global index of pronoun omission attestation. On the one hand, we have traditional and dialect-contact L1s, with a percentage of attestation of pronoun deletion of 40% and 37.5% respectively. The reader may recall from the previous discussion that these types included varieties that are/were not affected by language contact to a great extent and thus are the least likely to show simplification effects due to second-language acquisition and use. On the other hand, shift L1s, indigenized L2s, pidgin, and creoles exhibit higher attestation rates, ranging from 71.4% in the case of pidgins to 88.9% in indigenized L2 varieties. Figure 1 clearly reflects a split between the six variety types on the basis of the influence of second language acquisition/use, with primarily-L1 varieties (i.e., traditional and dialect-contact L1s) behaving fairly differently from the rest.

The percentages of attestation of pronoun omission are much more similar across regions than across variety types (cf. Figure 2). With the exceptions of Asia, in which pronoun omission is attested in all the varieties, and the British Isles, with a much lower attestation rate, all the regions fall in the 66.7% to 80% range. Pronoun omission is thus very frequently attested in most world regions, and especially in Asia, in line with Schröter and Kortmann's (2016) findings.

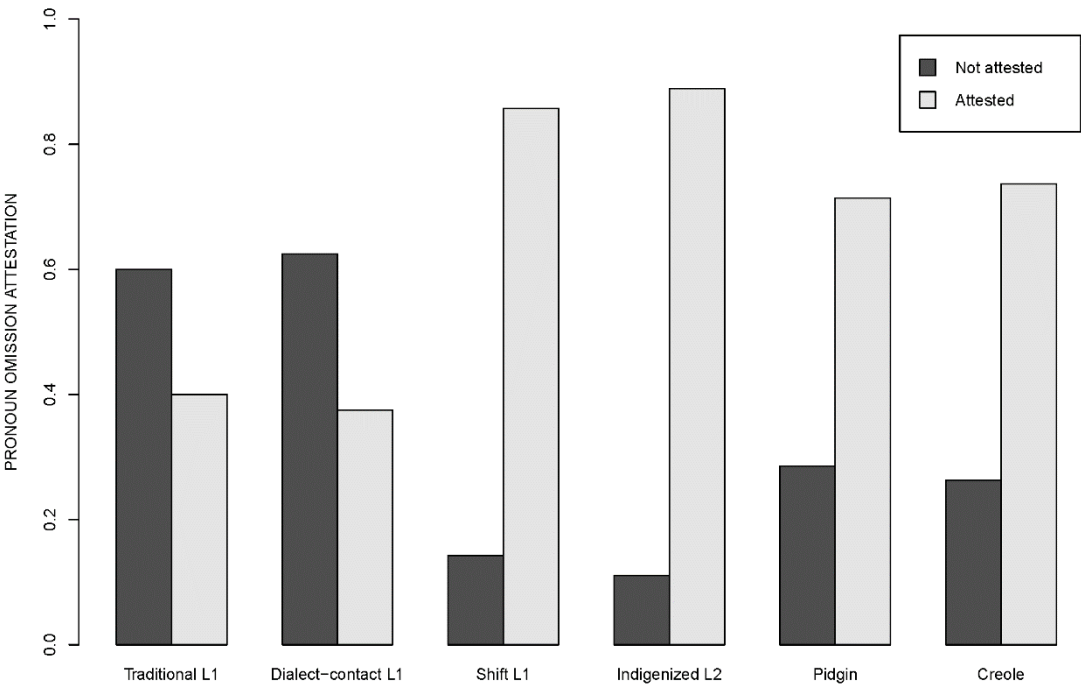


Figure 1. Pronoun omission attestation across variety types

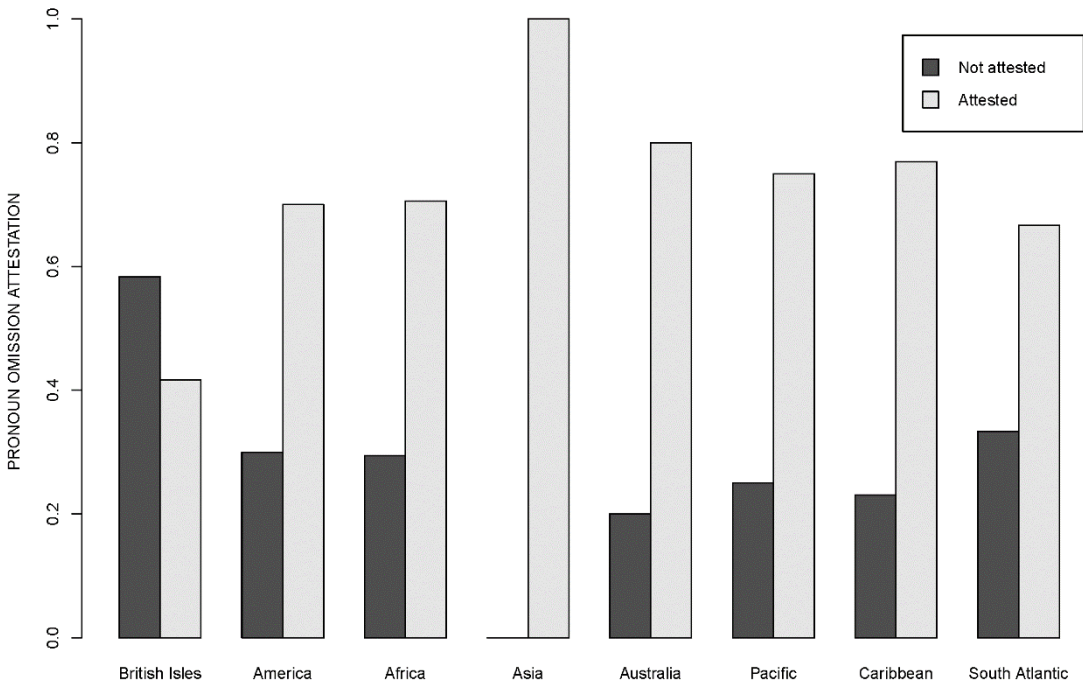


Figure 2. Pronoun omission attestation across regions

When the two factors, region and variety type, are included in a multifactorial analysis with pronoun omission attestation as a dependent variable, only the latter emerges as having a statistically significant effect (cf. Table 6). After removing the non-significant predictor region from the model, both Traditional and Dialect-contact L1s exhibit a significantly lower probability of attestation of pronoun omission features than Shit L1s and Indigenized L2s, and a marginally significant lower probability than Creoles (cf. Tables 29-30 in the Appendix for further detail). Pidgins, despite behaving like Creoles (cf. Figure 1), do not differ significantly from any of the other variety types. The findings suggest a division between varieties affected by second-language acquisition and those that are/were not and thus reinforce the previous interpretation of the results in Figure 1 above.

Table 6. Single term deletions from a model containing both variety type and region as predictors of pronoun omission attestation

Predictors	D.f.	Deviance	AIC	Lik. Ratio Test	P-value
Variety type	5	80.98	96.98	11.86	< 0.05
Region	7	78.36	90.36	9.25	0.235

Plotting the residuals of the model with variety type as a predictor of pronoun omission attestation enable us to evaluate whether regional patterns can be identified after removing the influence of variety type. Figure 3 displays the predictions of the model in the vertical axis and residual deviances in the horizontal axis. Values over the 50% threshold in the vertical axis signal that a given variety is predicted to attest at least one pronoun omission feature, while those below the threshold indicate that a variety is predicted not to attest any pronoun omission features. In the horizontal axis, positive residuals imply that the prediction for a

given variety was too low and negative values that the prediction was too high. In addition, varieties are grouped with respect to their actual observed value for the dependent variable: varieties whose relative positions in the plot are represented by means of triangles are those which attest pronoun omission features, while varieties represented by means of circles are those whose observed value for pronoun omission attestation is not attested. The idea is to see whether region plays a residual role in the attestation of pronoun omission features after accounting for the influence of variety type.

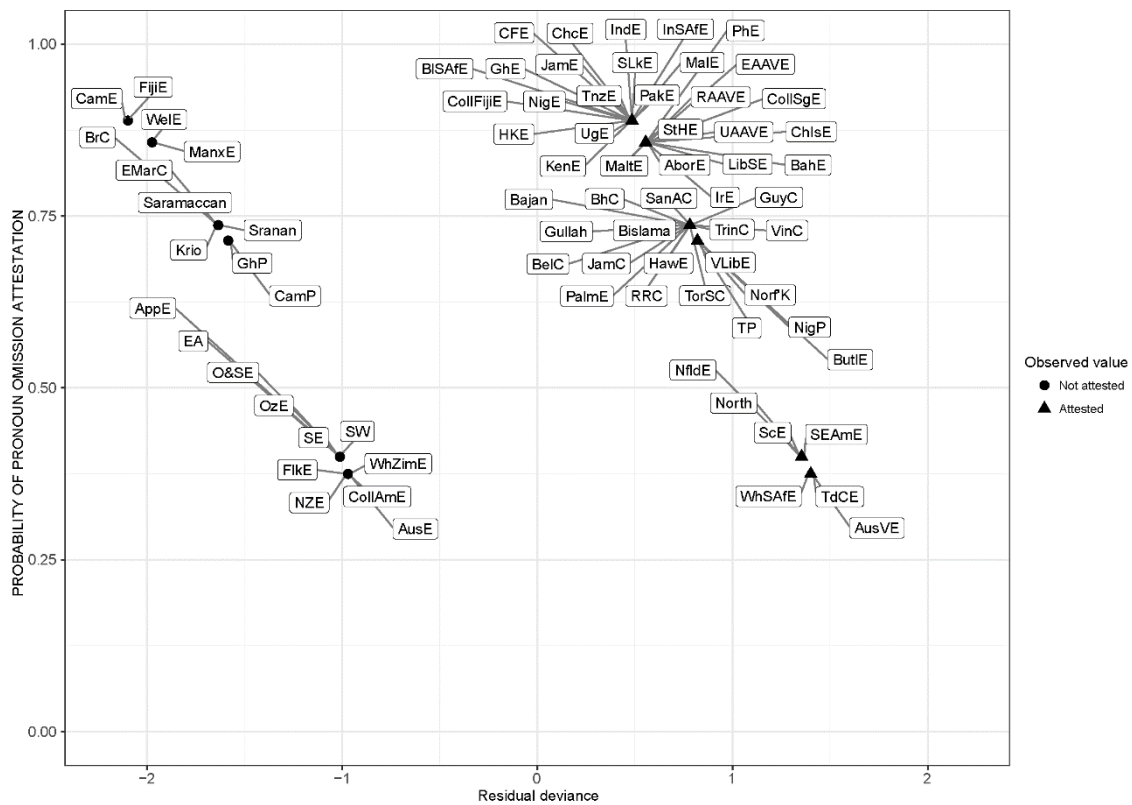


Figure 3. Residuals grouped by observed value

Four different clusters of varieties can be identified in Figure 3. In the top right corner of Figure 1 we find those varieties that attest pronoun omission features and are also predicted to do so by the model; therefore, they exhibit a low residual deviance. Similarly, the behaviour of those varieties in the bottom left corner also fit the predictions of the model, as they do not

attest any pronoun omission features and they are correctly predicted to exhibit a probability of pronoun omission attestation that is lower than 50%. On the other hand, the varieties in the bottom right and top left corners of Figure 2 do not behave as predicted by the model: either they are predicted to attest pronoun omission features but do not do so (top left corner) or, vice versa, they are predicted not to attest pronoun omission features but they do so (bottom left corner). It could be that these varieties which are not correctly classified by the model exhibited some residual regional effects, but it does not seem to be the case. There are some varieties which the model fails to classify whose behaviour could indeed be explained on the basis of their region: in the top left corner of Figure 3, BrC, a creole, and ManxE and WeE, Shift L1 varieties, are not surprisingly predicted to attest pronoun omission features but they do not do so according to *eWAVE*, possibly because they are located in the British Isles and British Isles varieties disfavour the attestation of pronouns omission features (cf. Figure 2). Similarly, NfldE, AusVE, TdCE, and WhSAfE in the bottom right corner are predicted not to attest pronoun omission features because they are Traditional and Dialect-contact L1s (cf. Figure 1), but they in fact attest pronoun omission features, which could be due to the fact that they are spoken in regions which favour attestation (America, Australia, South Atlantic, and Africa, respectively). However, some varieties are not correctly classified by the model but their observed values cannot be explained in terms of the world region in which they are located; this is the case of North and ScE in the bottom right corner and FijiE, EMarC, Saramaccan, Sranan, Krio, GhP, CamE, and CamP in the top left corner of Figure 3. On the one hand, the varieties in the bottom right corner are predicted not to attest any pronoun omission features (because they are Traditional L1s) but they do so even though they belong to the British Isles. The varieties in the top left corner, on the other hand, are predicted to attest pronoun omission features because they are Indigenized L2s, Pidgins, and Creoles but

they do not in fact do so even though are spoken in the Pacific, the Caribbean, and Africa, all regions which are prone to pronoun omission (cf. Figure 2). It seems, therefore, that regional patterns are only residual even after accounting for the effect of variety type, and we can safely conclude that this is the most important predictor of pronoun omission attestation.

As regards the second global index, pronoun omission pervasiveness, which quantifies how pervasively pronoun deletion occurs in those varieties in which it is attested, the six variety types distinguished do not differ greatly in this respect. Figure 4 displays the median pervasiveness scores of each variety type, which range from 0.6 in the case of traditional L1s to 1.65 in creole varieties.

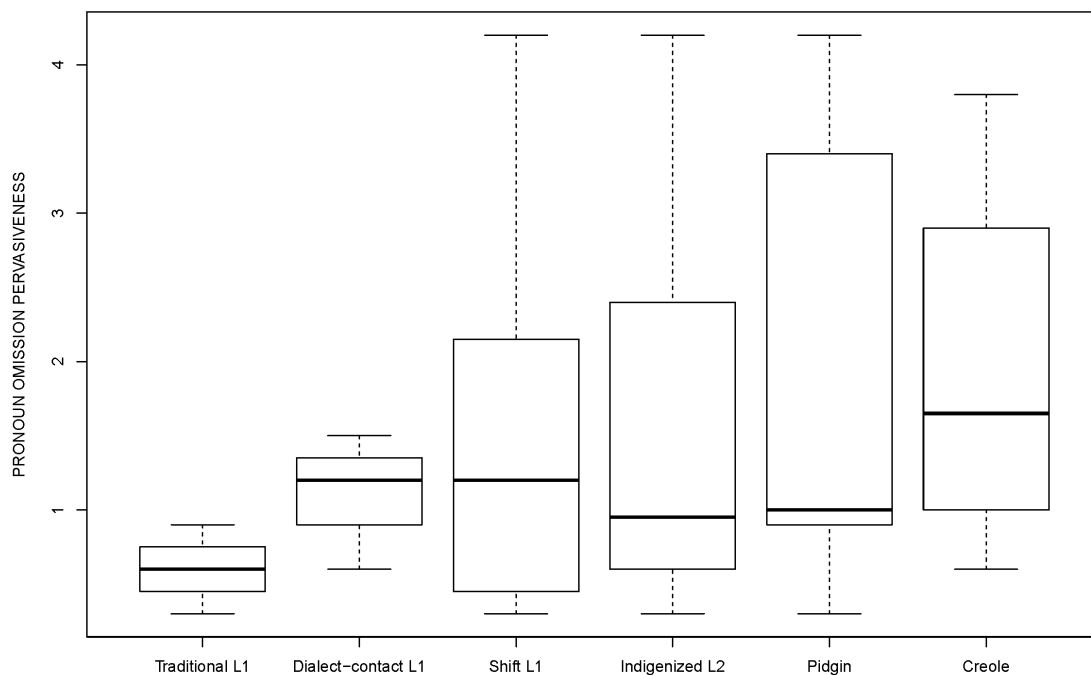


Figure 4. Pronoun omission pervasiveness across variety types

Region, on the other hand, displays a more interesting distribution. Figure 5 shows that pronoun deletion is particularly pervasive in Asia and the Pacific, and much less frequent in the British Isles, America, and Africa, with Australia, the Caribbean, and the South Atlantic

occupying an intermediate position. Interestingly, Figure 5 reveals the existence of two outliers, that is, varieties that behave very differently in comparison with the others in their regions. In the British Isles, MaltE, with a pervasiveness value of 3.5, differs widely from the median of the region of 0.6. This is not surprising if we consider the geographical location of Malta, which is not really in the British Isles at all but is included in that region for practical purposes, because it is the only European non-British variety in the sample (Kortmann and Lunkenheimer 2012, 3). Gullah, with a pervasiveness of 2.9, is the outlier in America, probably because it is the only creole in that region and thus differs in type from all the other American varieties.

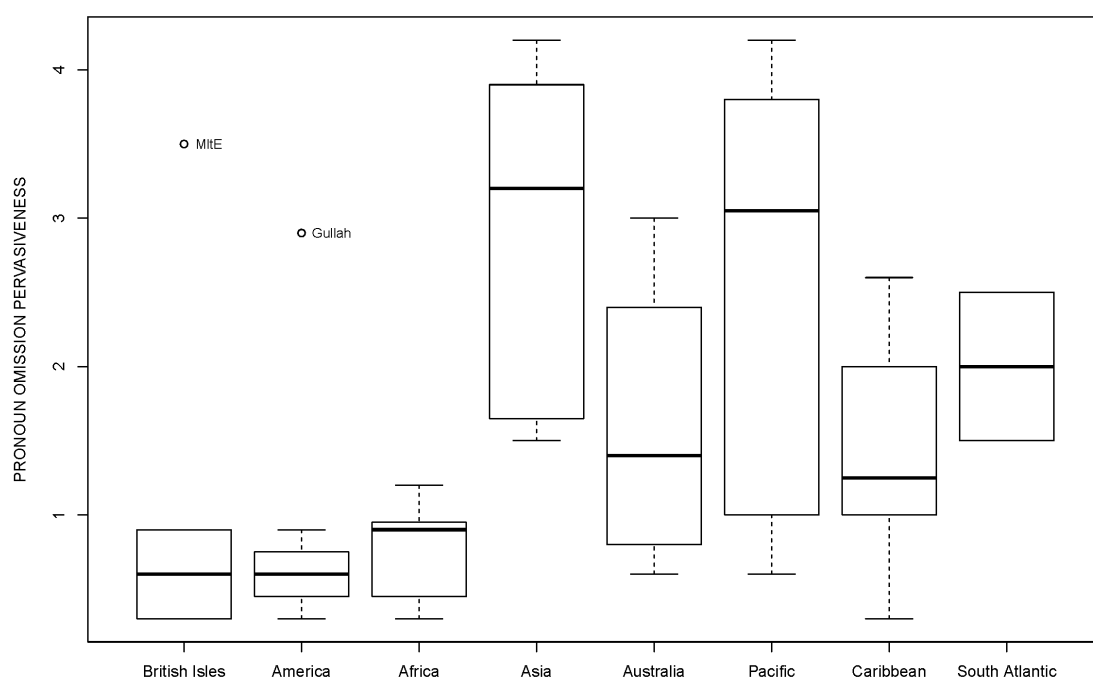


Figure 5. Pronoun omission pervasiveness across regions

The results of a linear regression model with pronoun omission pervasiveness as dependent variable and both variety type and region and predictors reveal that, not

surprisingly, only the latter has a significant effect (cf. Table 7). After removing variety type from the model, the following two groups of regions can be distinguished (cf. Tables 31-32 in the Appendix for further detail): British Isles, America, Africa, Australia and the Caribbean display a significant lower degree of pervasiveness of pronoun omission features than Asia and the Pacific. The latter two world regions, therefore, contain the varieties in which pronoun omission features are most pervasive.

Table 7. Single term deletions from a model containing both variety type and region as predictors of pronoun omission pervasiveness

Predictors	D.f.	Sum sq.	Res. sum sq.	AIC	F-value	P-value
Variety type	5	3.51	41.77	2.13	0.75	0.589
Region	7	30.14	68.40	24.76	4.61	< 0.001

Figure 6 displays the residuals of the linear regression model grouped by variety type (by means of different colours). The vertical axis shown the predictions of the model on the basis of region, and the horizontal axis the residual deviance. The position of the varieties in the plot is marked by numbers which represent their actual observed values. There are some varieties whose predicted values differ greatly from their observed values and which can be explained in terms of their variety types: GuyC, RRC, Gullah, and MalE exhibit higher observed pronoun omission pervasiveness values than predicted on the basis of the world regions in which they are spoken (namely, the Caribbean, Australia, America, and the British Isles, respectively), which could be due to the fact that the variety types to which they belong, namely Creoles and Shift L1s, are the ones with the highest median pronoun omission pervasiveness values (cf. Figure 4). Similarly, the Pacific Pidgin TP, and Asian Indigenized L2s PakE and SlkE display higher predicted values than observed, and this could in part be

explained on the basis of their variety types, since Pidgins and Indigenized L2s do not exhibit particularly high median pronoun omission pervasiveness values (cf. Figure 4). However, there are also varieties whose misclassification in the model cannot be accounted for in similar terms. HawE, for instance, has a much higher predicted pronoun omission pervasiveness value than observed (despite being a Creole), and this is also the case for other varieties such as PhE or Bislama. Therefore, variety type does not emerge as an important predictor of pronoun omission pervasiveness even if we control for the effect of region.

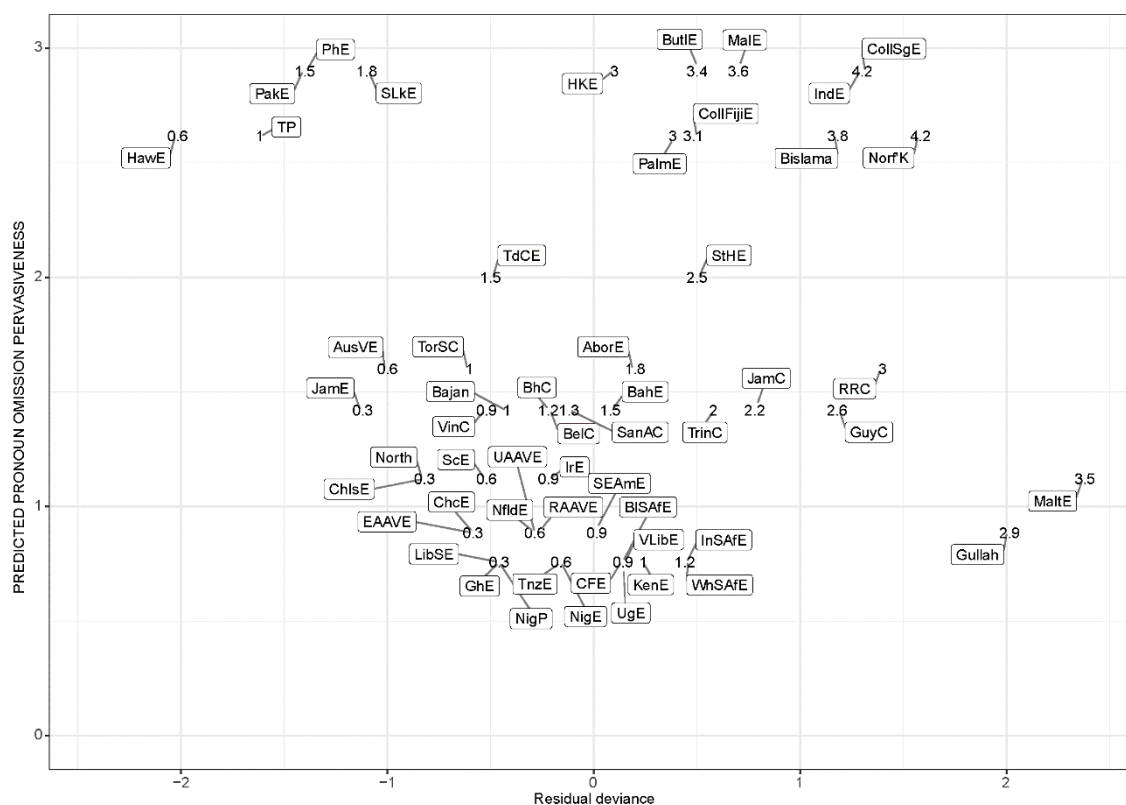


Figure 6. Residuals and observed values

To sum up the findings for the global indexes of pronoun omission attestation and pervasiveness, variety type influences the attestation of pronoun deletion, with those varieties affected by second-language acquisition and use exhibiting higher attestation rates than

primarily-L1 varieties. Region, on the other hand, accounts for the pervasiveness of pronoun omission in the varieties in which it is attested, with Asian and Pacific varieties outperforming the rest. The following section deals with the second set of results, that is, the attestation and pervasiveness of individual pronoun deletion features.

4.3.2 Attestation and pervasiveness of individual features

At the level of individual features, only region emerges as significant. Out of the five features in *eWAVE*, the attestation rates of object (F42), non-referential subject (F44), referential *it* (F46), and non-referential *it* omission (F47) are significantly higher in some regions, especially Asia and the Pacific. In addition, region accounts for the pervasiveness of object (F42) and non-referential *it* omission (F47). For the sake of brevity, only significant results are discussed in the remainder of this section (cf. Tables 33-44 in the Appendix for model statistics).

First, object omission is more frequently attested in Asia and the Pacific than in the British Isles, America, and the Caribbean, with Africa, Australia, and the South Atlantic occupying an intermediate position (cf. Figure 7): all the varieties in Asia and 75% of those spoken in the Pacific allow the omission of objects, while only 8.3%, 10%, and 15.4% do so in the British Isles, America, and the Caribbean respectively. Similarly, as shown in Figure 8, the pervasiveness of object omission in those varieties in which it is attested is higher in Asia and Australia, with a median of 0.8, and the Pacific, with a median value of 0.6.³⁹ Africa, on the other hand, exhibits the lowest median object omission pervasiveness (0.3).

³⁹ The British Isles exhibit the highest median object omission pervasiveness of all the regions, but this is based on only one variety, namely MaltE, which is in fact included in the British Isle region for convenience (cf. Section 4.3.1). Similarly, the median value of America is based on only one variety, the creole Gullah, which is not representative of its region. Finally, the

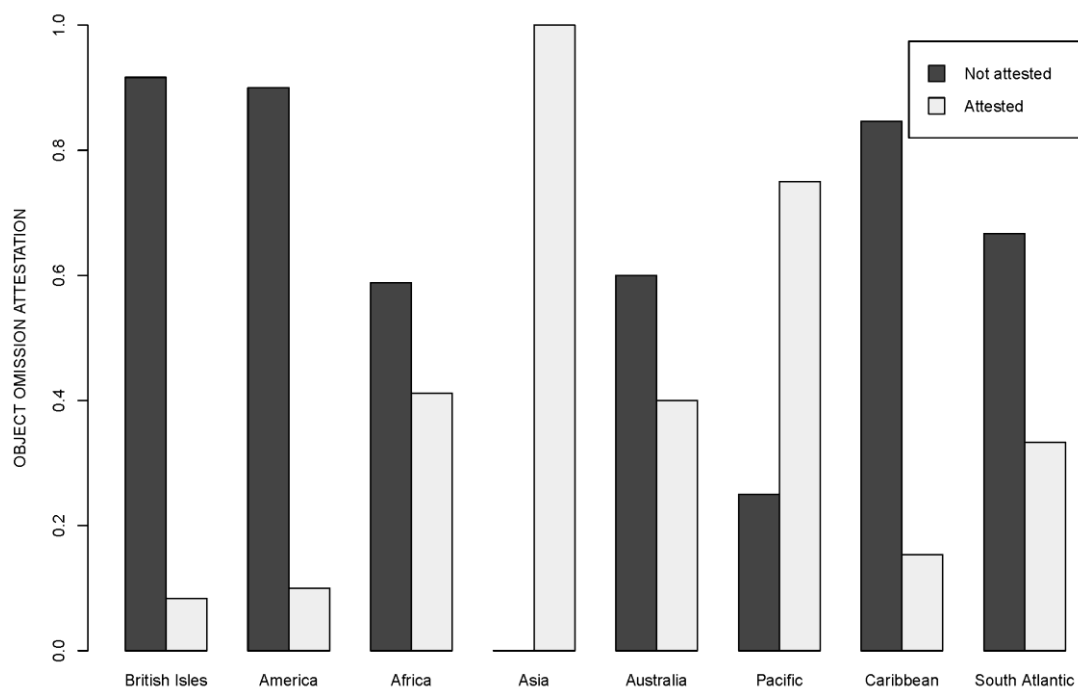


Figure 7. Object omission attestation across regions

As regards significant differences between the regions, the probability of attesting object omission is higher in the Pacific than in the British Isles, America, and the Caribbean. Moreover, Africa exhibits a marginally significantly higher probability than the British Isles. Therefore, there is a continuum of regions as to their degree of attestation of object omission, with the British Isles (low attestation rate of object omission) in one end and the Pacific (high attestation rate) in the other, with the rest of the regions occupying intermediate positions. With respect to the pervasiveness of omitted objects across regions, Asia, Australia, and the Pacific differ significantly from the Caribbean (and marginally significantly from America, but cf. footnote 39). In addition, Asia and the Pacific exhibit higher predicted object omission pervasiveness values than Africa: the contrast between Asia and Africa reaches significance, median object omission pervasiveness displayed by the Caribbean and the South Atlantic regions are based on only two and one varieties, respectively, and should, therefore, be taken with care.

while that between the Pacific and Africa is only marginally significant. The British Isles are also involved in statistically significant contrasts with America, Africa, and the Caribbean (but cf. footnote 39). It seems, therefore, that we have two more or less coherent groups of regions: Asia, Australia, and the Pacific, on the one hand, with high object omission pervasiveness values, and Africa and the Caribbean, on the other, with much lower values.

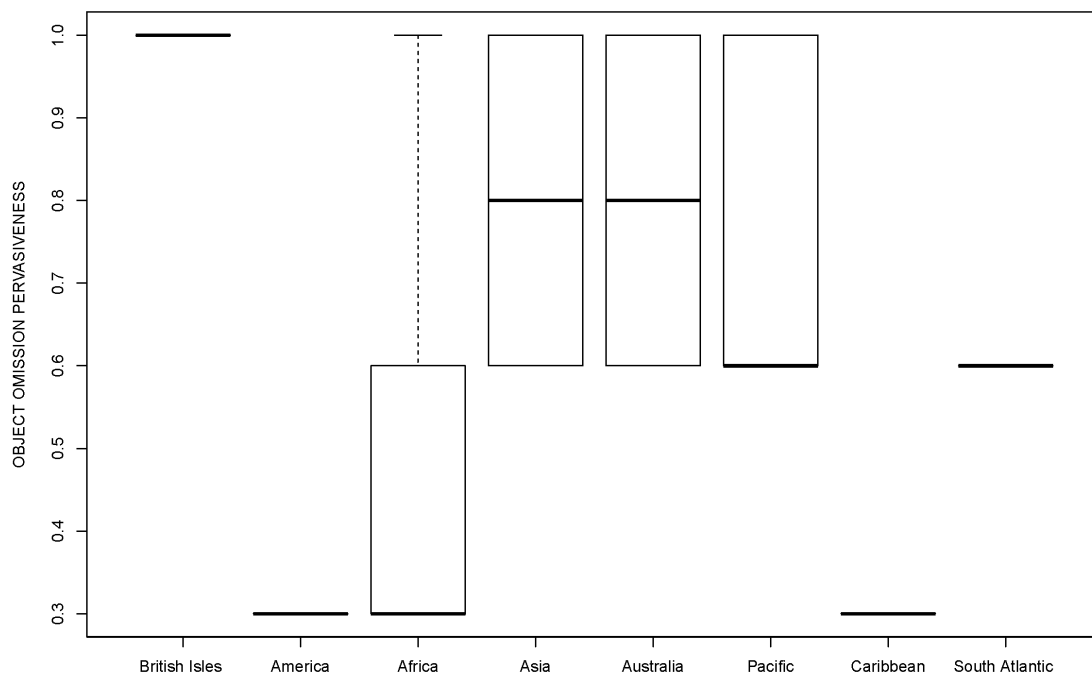


Figure 8. Object omission pervasiveness across regions

A similar picture emerges with respect to omitted non-referential subjects, as seen in Figure 9: as in the case of omitted objects, Asian and Pacific varieties exhibit higher attestation rates of non-referential subject omission, to which we can add the South Atlantic and Australian regions. Contrariwise, the British Isles, America, Africa, and the Caribbean display lower percentages of attestation, falling in the 15.4% to 30% range. Significant differences emerge between Asia, Australia, and the Pacific, on the one hand, with high

probabilities of non-referential subject omission attestation, and the British Isles, Africa, and the Caribbean, on the other, with lower values (the contrasts between Australia and the latter three regions are only marginally significant). Furthermore, Asia differs significantly from America.

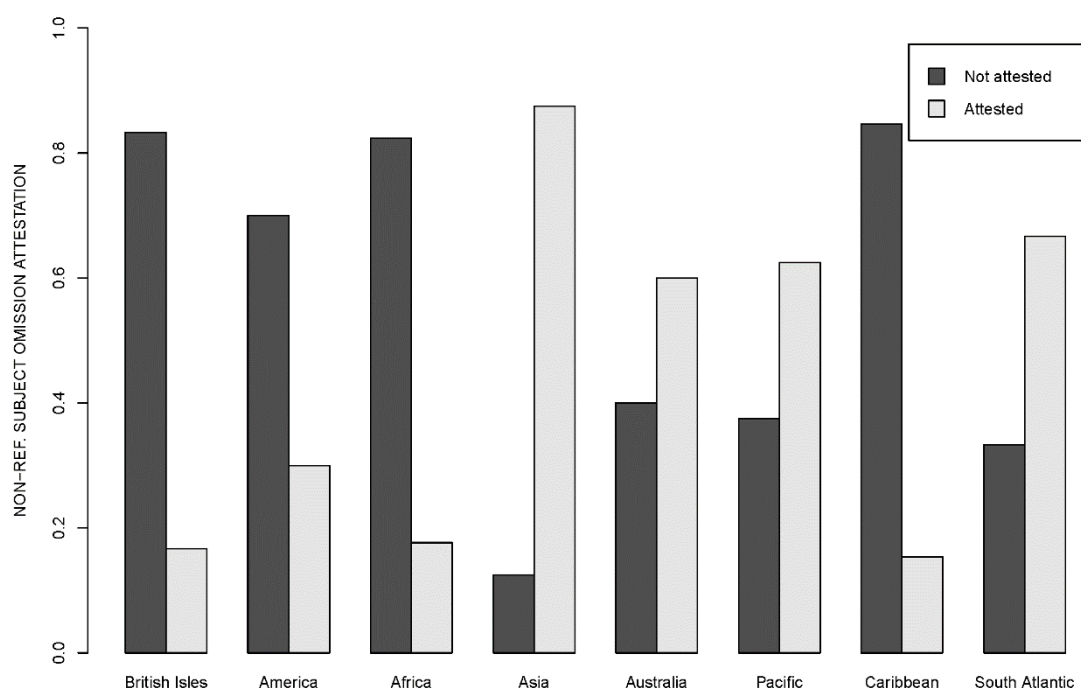


Figure 9. Non-referential subject omission attestation across regions

The attestation rate of referential *it* omission in *it is*-constructions also differs across regions (cf. Figure 10). Again, Asia and the Pacific can be found among the regions where omission is more frequently attested, with a percentage of referential *it* omission attestation of 50% in both areas. In this case, however, it is the Caribbean which displays the highest attestation rate (76.9%), followed by South Atlantic varieties (66.7%). American varieties, with an attestation rate of 40%, occupy an intermediate position, while Africa, the British Isles, and Australia are located at the lowest end of the continuum, with 29.4%, 8.3%, and 0%

attestation respectively. Statistically significant differences arise between the Caribbean, on the one hand, and Africa, the British Isles, and America. Similarly, Asia and the Pacific differ marginally significantly from the British Isles. Referential *it* omission, then, seems to be characteristic of the Caribbean and, to a certain extent, Asia and the Pacific.

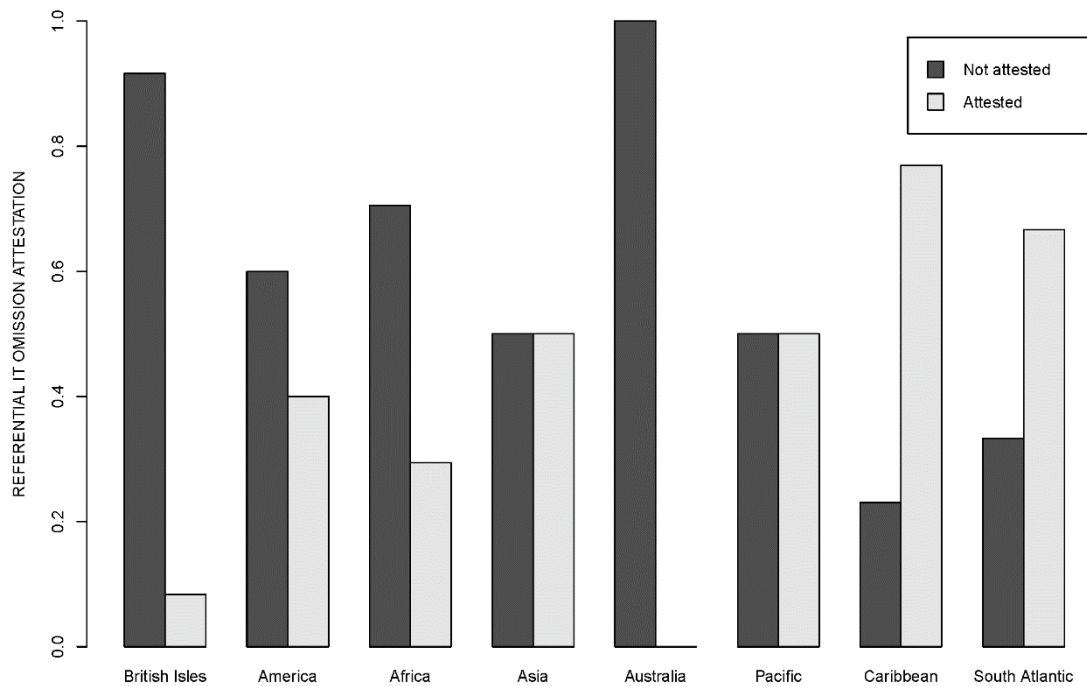


Figure 10. Attestation of referential *it* omission across regions

Finally, the deletion of *it* in non-referential *it is*-constructions, as shown in Figure 11, is only relatively frequently attested in Asia (62.5%), the Pacific (50%), and the South Atlantic (66.7%). The other regions exhibit attestation rates in the 8.3% to 30.8% range. With respect to its pervasiveness (cf. Figure 12), non-referential *it* omission displays higher values in Asia

(0.6), the Pacific (0.8), and the Caribbean (1) than in the other world regions.⁴⁰ Interestingly, HKE emerges as an outlier in Asia, since its non-referential *it* omission pervasiveness value of 1 is higher than the median of the region. This result agrees with Schröter and Kortmann (2016, 236), who also found HKE to exhibit a high frequency of omission in non-referential *it* *is*-constructions. As regards significant differences between the regions, Asia and the Pacific display a higher probability of non-referential *it* omission attestation than the British Isles and America (the contrasts are only marginally significant in the case of the Pacific). Furthermore, Asia differ marginally significantly from Africa. Again, Asia and the Pacific emerge as the pronoun omission regions par excellence. As to the pervasiveness of non-referential *it* omission, Asia, the Pacific, and the Caribbean exhibit a higher value than Africa. In addition, the Pacific and the Caribbean differ marginally significantly from the British Isles and America (but cf. footnote 40). The deletion of non-referential *it* in *it is*-constructions, therefore, is only frequently attested in Asia and the Pacific and, in those varieties in which it is attested, it is particularly pervasive in the Caribbean, the Pacific, and, to a certain extent, Asia.

⁴⁰ As in the case of Figure 8, the values exhibited by the British Isles and America are problematic since they are based on only one variety: MaltE in the British Isles and SEAmE in America. Australia is not shown in Figure 12 because non-referential *it* omission is not attested in any Australian varieties (cf. Figure 11).

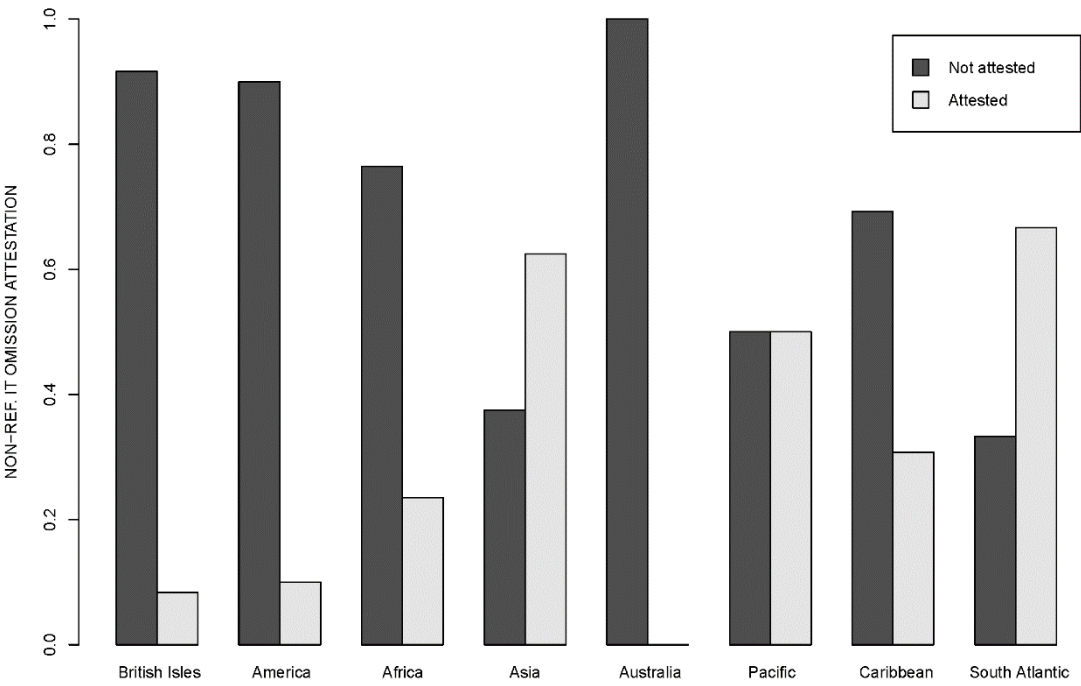


Figure 11. Attestation of non-referential *it* omission across regions

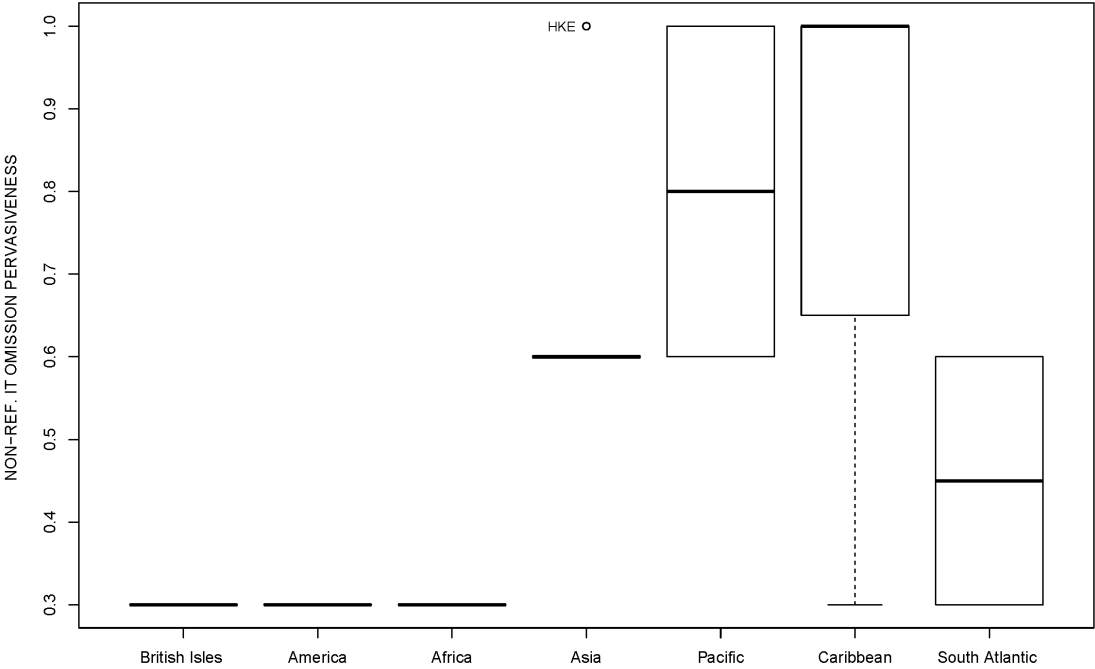


Figure 12. Non-referential *it* omission pervasiveness across regions

The evidence summarized in this section indicates that, at the level of some individual pronoun omission features, it is the region in which the varieties are spoken rather than their type which accounts for the attestation rates and pervasiveness values encountered. Asia and the Pacific are again the regions where pronoun deletion is more frequently attested and more pervasive: object, non-referential subject, referential *it*, and non-referential *it* deletion are commonly found in these geographical areas, with object and non-referential *it* omission also being particularly pervasive. In addition, Australia and the Caribbean exhibit higher attestation rates and pervasiveness values than other regions in particular contexts: with object pronouns (pervasiveness) and non-referential subject pronouns (attestation) in the case of Australia, and with referential *it* (attestation) and non-referential *it* (pervasiveness) in *it is*-constructions in the Caribbean. These results, as argued in the next section, suggest that substrate effects should figure prominently in any account of the occurrence of pronoun omission in varieties of English, but that the role of variety type must not be underestimated.

4.4 SIMPLIFICATION AND SUBSTRATE EFFECTS REVISITED

The findings presented in Section 4.3 clearly indicate that we need both variety type, as a proxy for degree of contact and second-language acquisition effects, and region, as a proxy for substrate influence, to explain the distribution of pronoun omission features in varieties of English. Variety type, as operationalized here, reflected a split between primarily-L1 (i.e., traditional and dialect-contact L1s) and L2-influenced varieties (i.e., shift L1s, indigenized L2s, pidgins, and creoles), with the latter being hypothesized to be more prone to simplification processes due to the impact of second-language acquisition and use. Pronoun omission, as argued in Chapter 3, results in economical and efficient structures and thus should be favoured in the L2 group. This hypothesized split between the six variety types was

also noticeable in the results of the study and verified by the statistical analysis of the data: pronoun omission was more frequently attested in the L2-influenced varieties than in the two primarily-L1 types. Furthermore, no significant differences were found among the L2-influenced types, which means that they behaved similarly with respect to the rate of attestation of pronoun omission. This distribution agrees with Kortmann (2013), who found that type accounted for the varieties' overall morphosyntactic profiles, understood as the co-presence and co-absence of features. At a smaller scale, type can also be claimed to account for the varieties' pronoun omission profiles, that is, the co-presence and co-absence of pronoun omission features, with higher attestation rates in shift L1s, indigenized L2s, pidgins, and creoles.

Geographical patterns, on the other hand, are considered here to result from the influence of substrate languages. Despite not having a significant influence on the overall attestation rate of pronoun deletion, regional effects were visible on its global pervasiveness in those varieties in which it was attested: Asia and the Pacific exhibited much higher median pervasiveness values than the other varieties, especially the British Isles, America, and Africa, in which pronoun omission, despite being attested, was extremely rare. A very similar distribution is found in *WALS*, as shown in Figure 13: most languages in which subject pronouns are obligatory are located in Africa, America, and Northern Europe, while those with optional subject pronouns are mostly found in South and Southeast Asia, the Pacific, and, to a certain extent, Australia. It seems reasonable to suggest that the presence of pronoun deletion languages in the varieties' linguistic substrate favours the omission of pronouns, thus increasing their frequency of occurrence in the speakers' second language.⁴¹ This conclusion

⁴¹ However, the vast majority of languages (437) included in *WALS* in the sample for F101A "Expression of pronominal subjects" (Dryer 2013) belong to a different category, namely those in which subject pronouns are expressed by means of

is consistent with Kortmann's (2013) first scenario for the influence of geography on variation in the Anglophone world. He hypothesizes that region may explain the degree of pervasiveness of features in a specific group of varieties, and this is exactly what the findings for pronoun omission seem to reflect.

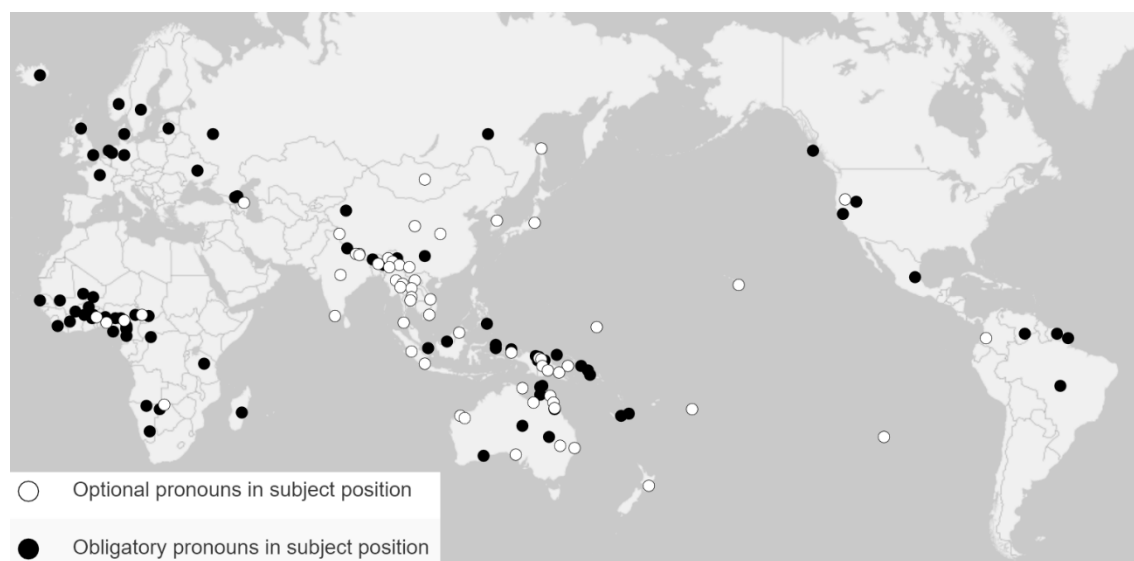


Figure 13. Expression of pronominal subjects in WALS (F101A)

Additional regional, and thus substrate, effects were found at the level of individual features. The attestation rates of object, non-referential subject, referential *it*, and non-referential *it* omission exhibited regional distributions, with the four deletion features being commonly attested in Asia and the Pacific. Moreover, Australian varieties displayed a high attestation rate of non-referential omitted subjects, and Caribbean varieties of referential *it* deletion in *it is*-constructions. The pervasiveness values of two features also seemed to reflect

affixes attached to verbs (i.e., subject-verb agreement languages). These languages, besides being numerous, are evenly distributed around the world, which poses a major problem for the substrate influence hypothesis: there is no a priori reason why subject-verb agreement languages which allow pronoun deletion should not favour omission in the relevant varieties. For instance, Spanish learners of English tend to transfer their pronoun omission patterns to the second language, and Spanish is a subject-verb agreement language (Filipović and Hawkins 2013, 152).

an influence of the region in which the varieties are spoken: object omission was more pervasive in Asia, Australia, and the Pacific and non-referential *it* omission in Asia, the Pacific, and the Caribbean. These results can be connected to Kortmann's (2013) third hypothetical scenario for regional effects, whereby the choice of specific constructions for the same grammatical function across varieties and the existence of local syntactic, semantic, and discourse constraints on features are explained by reference to the relevant substrate languages. The fact that omitted pronouns are found more frequently in specific contexts in certain regions may be connected to the substrate languages spoken in those regions. However, further research is needed in order to clarify this issue, since a complete account of the manifestation of substrate effects in varieties of English is beyond the scope of the present monograph.

The study presented in this chapter focused on (i) exploring the distribution of pronoun omission features in different variety types and world regions and (ii) untangling the independent effects of simplification and substrate influence. Asia and the Pacific emerged as the pronoun omission regions par excellence. The four individual features that exhibited regional distributions, namely object, non-referential subject, referential *it*, and non-referential *it* omission, were more commonly attested in Asian and Pacific varieties, and the overall pervasiveness of pronoun deletion features, that is, how frequently they occurred in actual language use in those varieties in which they were attested, as well as the pervasiveness of object and non-referential *it* omission, were also higher in these two regions. Varieties spoken in Asia and the Pacific thus show a stronger tendency to omit pronouns than those located in other geographical areas, which may have to do with their substrate languages. It was variety type, however, which accounted for the overall attestation of pronoun omission features: L2-influenced varieties were prone to the omission of pronouns, which may be explained as a

consequence of simplification processes due to second-language acquisition and use. Therefore, an explanation of the occurrence of pronoun omission on the basis of only one factor runs the risk of missing important insights on this linguistic phenomenon. The findings discussed here make evident that both variety type and region are necessary to reach a fuller understanding of the distribution of pronoun deletion features in varieties of English, since they seem to apply at different levels: type, and thus the varieties' socio-historical conditions, predicts the overall attestation rate of pronoun omission, while region, and thus influence from substrate languages, accounts for its pervasiveness in the varieties in which it is attested and for the attestation and pervasiveness of some individual features.

The present chapter has provided a bird's-eye perspective on the distribution of pronoun omission features in the Anglophone world. Despite being successful in uncovering the explanatory power of variety type and geography as predictors of pronoun omission in varieties of English, such a study cannot offer insights on fine-grained distributional patterns in actual language use. This is the goal of the next chapter, which focuses on referential subject pronoun deletion in three varieties of English, namely BrE, SgE, and IndE.

5 COMPLEXITY VARIANCE IN ENGLISH: PRONOUN OMISSION AND LANGUAGE CONTACT

The occurrence of pronoun omission features in varieties of English, as shown in Chapter 4, is clearly favoured by language contact, especially if the substrate language(s) also allow the possibility of deleting pronouns. However, the evidence presented in the previous chapter must necessarily be complemented by a more detailed analysis of the distribution of pronoun omission to examine under which conditions speakers delete pronouns. This is the aim of the present chapter.

Pronoun omission, as argued in Chapter 3, Section 3.4, simplifies structures by making them more economical. The grammar, on the other hand, becomes more complex if there is an extra variant, namely omitted pronouns, in the set of referential expressions, which does not happen if pronouns are invariably overtly expressed. In addition, the existence of this extra referential device may result in a larger set of rules or constraints regulating the choice between the different variants. These two dimensions of the complexity of pronoun omission are reflected in the distinction between structural and system complexity discussed in Chapter 2, Section 2.3.5, which is the focus of the present chapter. The main goal here is to examine the linguistic conditioning of omitted referential subject pronouns, that is, the structural contexts in which they are (dis)favoured by speakers and the importance of these contexts in

the choice between omitted and overt pronouns. The varieties examined are BrE, IndE, and SgE, which, as discussed in Section 5.2.2 below, were chosen because they instantiate different degrees of contact, variety types, and linguistic ecologies.

This chapter is divided into five sections. Section 5.1 deals with the purpose and aims of the study, as well as with the general hypotheses that are then tested in subsequent sections. Section 5.2 focuses on the data and methodology of the study, that is, on the corpus from which the data were extracted, the retrieval process, the operationalization of the variables examined, and the statistical analysis. In Section 5.3, the distribution of overt and omitted subject pronouns is univariately explored, while the conjoined effect of the variables on the choice between the two variants is taken up in Section 5.4. Finally, Section 5.5 concludes with some reflections about the relation between language contact and complexity.

5.1 AIMS AND PURPOSE OF THE STUDY

Language contact generally results in simplification: there is ample evidence of the simplifying effect of contact and many examples have been put forward in the specialized literature (cf. Chapter 2, Section 2.2.1). Much depends, however, on our definition of complexity. In Chapter 2, Section 2.3.5, three important distinctions were introduced: absolute versus relative complexity, structural versus system complexity, and overt versus hidden complexity. The first distinction relates to the perspective from which we approach the study of complexity. In brief, absolute approaches measure complexity objectively as the number of parts in a system, while relative approaches measure complexity as a function of how difficult or costly a language (or some part thereof) is to process and/or learn. The reader may recall that the approach followed in the present work is relative since it focuses on the preferences of language users. The second distinction pertains to the nature of the objects to

be measured: whereas system complexity metrics concentrate on the paradigms and rules mapping meanings to forms, structural metrics measure the outputs of those rules. In addition, within system complexity a further split is made between tectogrammatics and phenogrammatics, with the former referring to the number of different structural choices in a grammar and the latter to the number of rules that constrain the use of those choices. Finally, the overt versus hidden complexity distinction involves the type of linguistic processes considered: overt complexity refers to the explicit marking of grammatical categories and rules by means of morphosyntactic structures, while hidden complexity refers to the processes of pragmatic inference that are necessary for the correct interpretation of ambiguous and/or vague structures.

All these considerations were taken into account when the complexity of pronoun omission was assessed in Chapter 3, Section 3.4. From a structural perspective, pronoun omission results in economical and efficient structures that provide the hearer with early access to their ultimate syntactic and semantic representation. However, from a systemic point of view, it increases the complexity of the grammar. This applies both tectogrammatically, due to the existence of an extra referential device, and phenogrammatically, because extra rules or constraints are needed to explain variation between omitted and overt pronouns. Economical structures in turn entail less overt complexity, due to the decrease in the number of overt forms that have to be articulated and processed, but their lesser explicitness may result in more hidden complexity: in structures with omitted pronouns, there are fewer overt cues guiding the addressee to the correct referent, which implies additional inferential processes. This increase in hidden complexity, however, might be mitigated if the antecedent is easy to retrieve by virtue of being a highly accessible entity or event in the discourse.

Against this backdrop, the present chapter aims at measuring the (relative) structural and system complexities of BrE, IndE, and SgE by means of a corpus study on the distribution of omitted and overt referential subject pronouns. In addition, although the focus is primarily on overt complexity, potential changes in hidden complexity arising as a result of the presence or absence of referential subject pronouns are also considered. The study focuses on the effects and discriminating power of several language-internal constraints (cf. Section 5.2.3 below) on the probability of omission in order to delineate the pronoun-deletion grammars of the three varieties at hand. In addition, the varieties represent different degrees of contact and variety types, which allows us to investigate the influence of these language-external factors on the likelihood of pronoun omission. First, SgE and IndE are characterized by a high degree of language contact, while BrE, despite not being an isolated variety, is affected by contact to a much lower degree. Second, IndE is spoken as a second language by the vast majority of its speakers (Sharma 2012, 523), while BrE is clearly a native variety. SgE, on the other hand, lies in between these two varieties: despite having been spoken chiefly as a second language in the past, it is now in the process of becoming the first language of a sizeable proportion of the Singaporean population (Leimgruber 2013, 9). Lastly, the high-contact varieties IndE and SgE do not coincide in the languages present in their linguistic substrates, so differences between them may have also emerged as a result of substrate influence.

The primary goal of this study is to estimate the structural and system complexities of BrE, IndE, and SgE with respect to pronoun omission in order to isolate the influence of language contact on complexity. To this purpose, I use the metrics introduced in Chapter 3, Section 3.4, which are operationalized on the basis of regression and random forests analyses in Section 5.2.4 below. The different constraints said to influence the choice between omitted and overt referential subject pronouns reported in Chapter 3, Sections 3.2 and 3.3, are

considered here: first, the distribution of omitted and overt pronouns is explored with respect to each factor separately to identify the environments in which each alternative is more frequent and, second, the conjoined effect and explanatory power of the constraints are examined so as to determine how complex the varieties' pronoun-deletion grammars are.

Three general hypotheses are put to the test in the present chapter. The first one is related to the overall influence of language contact on complexity: pronoun omission should be more strongly preferred by speakers of SgE and IndE than by BrE speakers. Assuming that deleted pronouns result in more economical structures, without a concomitant loss in clarity if antecedents are highly accessible to the addressee, they should be favoured in the high-contact varieties due to the influence of second-language acquisition and use. The second and third hypotheses zoom in on the specific predictions for structural complexity, on the one hand, and system complexity, on the other. From a structural perspective, pronoun omission should be preferred in 'omission-friendly' environments, that is, when it is favoured by independent processing or cognitive considerations, such as priming effects or high referent accessibility. Furthermore, the aforementioned hypothesized spread of pronoun omission in SgE and IndE should be restricted to those environments in which omission does not disrupt the ordinary course of online processing (cf. Section 5.2.3.3). From a systemic perspective, IndE, being the only 'pure' second-language variety in the set, should exhibit a simpler pronoun-deletion grammar than either SgE or BrE. In other words, fewer grammatical constraints should influence the choice between omitted and overt pronouns in IndE than in the first-language varieties, thus reflecting the acquisitional preferences of second-language speakers. These three general hypotheses are complemented in the next section by several more specific ones related to the factors included in the analysis, that is, the expected effect of each of the constraints considered here on the distribution of omitted and overt pronouns.

5.2 DATA AND METHODOLOGY

5.2.1 Corpus and data retrieval process

The data used in the present study were retrieved from the British (*ICE-GB*), Indian (*ICE-IND*), and Singaporean (*ICE-SIN*) components of the *International Corpus of English (ICE)*. *ICE* comprises several national components, each of which contains approximately one million words: 600,000 from speech and 400,000 from written language. Each component aims at being a faithful representation of how English is spoken and written in the different countries included. A useful feature of the *ICE* family of corpora is that all the national components follow the same design and scheme for grammatical annotation, which means that, despite their small size, they are highly comparable. Each component contains 500 texts of approximately 2,000 words each. The classification of texts included in *ICE* is displayed in Table 8, with the number of texts in each category shown in brackets:

Table 8. Classification of texts in *ICE*

SPOKEN (300)	Dialogues (180)	<i>Private (100)</i>	Face-to-face Conversations (90) Telephone Calls (10)
		<i>Public (80)</i>	Classroom Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Legal Cross-examinations (10) Business Transactions (10)
	Monologues	<i>Unscripted (70)</i>	Spontaneous Commentaries (20)

	(120)		Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		<i>Scripted (50)</i>	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)
WRITTEN (200)	Non-printed (50)	<i>Student Writing (20)</i>	Student Essays (10) Exam Scripts (10)
		<i>Letters (30)</i>	Social Letters (15) Business Letters (15)
	Printed (150)	<i>Academic Writing (40)</i>	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		<i>Popular Writing (40)</i>	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		<i>Reportage (20)</i>	Press News Reports (20)
		<i>Instructional Writing (20)</i>	Administrative Writing (10) Skills/hobbies (10)
		<i>Persuasive Writing (10)</i>	Press Editorials (10)
		<i>Creative Writing (20)</i>	Novels and Short Stories (20)

Table 9. Texts included in the subsets of *ICE-GB*, *ICE-IND*, and *ICE-SIN*

SPOKEN (20)	Informal (10)	Face-to-face Conversations (8) Telephone Calls (2)
	Formal (10)	Classroom Lessons (1) Broadcast Discussions (1) Broadcast Interviews (2) Parliamentary Debates (2) Legal Cross-examinations (2) Business Transactions (2)
WRITTEN (20)	Informal (10)	Social Letters (10)
	Formal (10)	Academic Writing (2) Popular Writing (2) Press News Reports (2) Instructional Writing (2) Press Editorials (2)

Due to the difficulty of retrieving all the relevant instances of omitted pronouns in the corpus, which had to be manually extracted, only a relatively small subset of the texts in *ICE-GB*, *ICE-IND*, and *ICE-SIN* was used. Forty texts were selected from each component, which amounted to 83,261 words from *ICE-GB*, 87,952 from *ICE-IND*, and 81,503 from *ICE-SIN*.⁴² So as to achieve a balanced sample, half of the texts selected from each component contained spoken language and the other half written language. Furthermore, within each of these mediums of production, half of the texts represented informal registers and the other half

⁴² Stretches of text marked as extra-corpus data (<X>...</X>) were excluded from the count, and so were instances of omitted and overt pronouns found in these materials.

formal ones: spoken informal texts were taken from ‘Face-to-face Conversations’ and ‘Telephone Calls;’ spoken formal texts were extracted from ‘Classroom Lessons,’ ‘Broadcast Discussions,’ ‘Broadcast Interviews,’ ‘Parliamentary Debates,’ ‘Legal Cross-examinations,’ and ‘Business Transactions;’ written informal texts were selected from the ‘Social Letter’ category; and, finally, written formal texts were taken from ‘Academic Writing,’ ‘Popular Writing,’ ‘Press News Reports,’ ‘Instructional Writing,’ and ‘Press Editorials.’ Table 9 summarizes the types of texts included in the subset used in the present study, with the number of texts in each category shown in brackets:

From this sample of *ICE*, omitted pronouns were manually retrieved by reading the texts and identifying the relevant observations. This process resulted in a total number of instances of omitted pronouns in the three varieties of 1,229. A random sample of 1,229 overt pronouns was then automatically retrieved using *WordSmith Tools* (Scott 2012). The following types of instances were not included in the final dataset of 2,458 observations analysed:

- Examples of omitted and overt pronouns in non-subject position were not taken into consideration in the present study (cf. example (36)). The reasons for this exclusion are that (i) the constraints influencing the choice between the two alternatives might vary across syntactic positions, and (ii) distinguishing between an omitted object and an intransitive use of a verb is not always straightforward.
- Similarly, non-referential subject pronouns (cf. example (37)) were not included since, as argued by Williams (1988), they display an idiosyncratic omission behaviour and, being semantically empty, cannot be analysed in terms of the accessibility of their referents, which is one of the factors examined in the present study.

- Cases of omission in which the following auxiliary verb was also dropped (cf. example (38)) were not considered here because the linguistic constraints that apply to this type of instances might be different from those applying to cases containing only an omitted subject (cf. Chapter 3, Section 3.3).
- Instances in imperative sentences (cf. example (39)) were excluded from the sample since subject omission seems to be the default alternative in these contexts (cf. Chapter 3, Section 3.3).
- Overt pronouns in fixed expressions, such as tag questions (cf. example (40)), were also disregarded as no variation was found in these contexts, that is, pronouns are always overtly expressed.
- Finally, repeated instances of overt pronouns (cf. example (41)) were eliminated from the final dataset so that each instance of an overt pronoun only appears once in the dataset.

(36) And **fish_i** is too expensive [...] He won't give **Ø_i** (ICE-IND:S1A-003#113-116)

(37) **Ø** Was a pleasant surprise to hear your voice again from the other end of the line.
(ICE-SIN:W1B-001#5:1)

(38) **Ø** Glad you could write and tell us how you were doing. (ICE-SIN:W1B-014#3)

(39) **Ø** send lots of love to Lee when you speak/see him and of course that goes for Simey also. (ICE-GB:W1B-005#38)

(40) Terror isn't it, sounds like Aristar Tony doesn't **it**? (ICE-SIN:W1B-005#7)

(41) Yeah well **I** I'll phone him up and work on him (ICE-GB:S1A-027#23)

5.2.2 Varieties selected

The materials from *ICE-GB* were included to act as a low-contact standard to which the other two varieties could be compared. However, the choice of IndE and SgE rather than any other high-contact variety merits an explanation. The differences in the status of English between India and Singapore enable us to examine the influence of language-external factors on the complexity of these varieties. English was introduced in India in the seventeenth century after it was colonized by the British. Nowadays, there are approximately a hundred million speakers of English in the country, of which only 250,000 are native speakers (Sharma 2012, 523). Most users, therefore, speak English as a second or third language, making IndE a non-native variety. English, together with Hindi, is one of the co-official languages of India, although its use is chiefly restricted to certain domains of life, such as the government, administration, politics, higher education, the legal system, business and the media. It also functions as an “interethnically neutral link language” (E. Schneider 2007, 167), but it is not a sign of Indian identity. The current status of English in India has been described as a “steady state” (Mukherjee 2007, 158) in which both progressive and conservative forces are at play. The most important progressive forces are the linguistic innovations that distinguish IndE from other varieties, but also certain developments allow us to entertain the possibility of English becoming more established in India in the future. Among these are the increase in the number of literary works written in English by Indian authors and the recent inclusion in the syllabus of a compulsory English subject in primary education. On the other hand, there are also conservative forces that hold back the spread of English in the country and make it difficult for IndE to become a marker of Indian identity. First, teachers of English follow predominantly a British norm, which hinders the establishment of an Indian standard. Second, and most importantly, many Indians consider the innovative features characteristic of IndE as

grammatical errors that must be avoided, and instead hold native varieties as the appropriate and correct ones. In terms of Trudgill's (2011) typology of contact situations mentioned in Chapter 2, Section 2.2.3, IndE is an example of the first type, that is, those characterized by short-term adult second-language acquisition.

India is a highly multicultural and multilingual nation. The Constitution recognizes eighteen official languages (Mukherjee 2007), but many more are spoken in the country. The majority of these languages belong either to the Indo-Aryan or to the Dravidian families. Hindi, which is part of the Indo-Aryan family, has a special status alongside English as the language used for official purposes. It is the native language of approximately 35% of the Indian population, which makes it the language with the largest number of speakers in India and, therefore, an important substrate language for IndE. In Hindi, pronoun omission is commonly used as a device to establish cohesion in oral and written language, but also in Dravidian languages such as Malayalam and in other South Asian languages (Moag and Poletto 1991). Omitted pronouns tend to refer to known or established antecedents in discourse, and they are the unmarked option in conversation (Kachru 2006, 258–59).

The status of English in Singapore deviates extensively from the one it has in India. Approximately 80% of the Singaporean population are literate in English and, besides being the language of politics, administration, the courts, and education, it is also used in more informal and intimate contexts: the percentage of Singaporeans using English as a dominant home language rose from 11.6% in 1980 to 18.8% in 1990, 23% in the year 2000, and 32.3% in 2010 (Leimgruber 2013, 9). The attitude towards the local variety differs widely from that of IndE speakers. SgE, with all its particular features, is not only accepted but also functions as an identity carrier. After the 1970s, Singapore became an industrialized and economically prosperous country with a blend of European and Asian influences (E. Schneider 2007, 155–

56). SgE came to be associated with this new cultural and national identity and, as a result, Singaporeans developed a positive attitude towards it. In terms of Trudgill's typology, SgE is moving from the first to the second type: while there are still many who do not speak it as a native language, childhood bilingualism is definitely on the increase.

The majority of the population of Singapore descend from the Chinese, Malay and Indonesian immigrants that moved to the island in the nineteenth century after its annexation by the British in 1819 (Bao 2001). At the beginning, people of Chinese and Malay descent were the major ethnic groups, and their proportions were very similar, but from the 1840s onwards the population of Chinese origin grew rapidly and became the largest one in the island. This distribution still applies today, and thus Chinese is a very important substrate language for SgE. The nineteenth-century Chinese immigrants that moved to Singapore spoke many different dialects, but since 1979 the government has advocated the use of the Mandarin variety among the Chinese population (E. Schneider 2007, 156). Today, it is one of the official languages of Singapore, together with Malay, Tamil and English. In Mandarin Chinese, entities which are understood from the context or which refer to general or nonspecific antecedents may be omitted (Li and Thompson 1989, 658–59). A common phenomenon in Chinese discourse is what Li and Thompson (1989, 659) call a “topic chain,” that is, when “a referent is referred to in the first clause, and then there follow several more clauses talking about the same referent but not overtly mentioning that referent.” Overall, overt pronouns are only used in those situations in which the speaker wants to emphasize or highlight an antecedent (Li and Thompson 1989, 663–64).

5.2.3 Variables included in the analysis

The final 2,458-instance dataset, retrieved from the three subsets of texts taken from *ICE-GB*, *ICE-IND*, and *ICE-SIN*, was then annotated for several language-internal and language-external variables. Out of all the factors influencing pronoun omission mentioned in Chapter 3, a total number of sixteen are examined in the present study. Since turn position, polarity, semantic class of the verb, and tense-aspect-mood of the verb were not found to influence the choice between overt and deleted subject pronouns in AmE by Torres Cacoullos and Travis (2014), these variables are not considered here. Let us now review the sixteen factors explored, starting with the language-external constraints.

5.2.3.1 Language-external constraints

Since one of the aims of the present study is to compare the pronoun-deletion grammars of the three varieties examined, a variable that must necessarily be included in the analysis is ‘Variety.’ The three levels considered are the varieties themselves, that is, BrE, IndE, and SgE. As mentioned above, pronoun omission is hypothesized to be more frequent in the high-contact varieties than in BrE due to the influence of second-language acquisition and use. Pronoun omission is also expected to have increased in use in these two varieties in those contexts in which it is not disfavoured by independent processing or cognitive considerations.

A further language-external constraint examined is ‘Text type’ so as to assess the effect of register and medium of production in the expression of referential subject pronouns. This variable comprises the four categories distinguished in the compilation of the subsets of texts from *ICE-GB*, *ICE-IND*, and *ICE-SIN*, that is, spoken informal, spoken formal, written informal, and written formal. There are potentially conflicting tendencies with respect to the influence of text type on subject omission in the present study. As mentioned in reference

grammars (cf. Chapter 3, Section 3.3), omission in English is common in coordinate clauses in all mediums and registers but, in initial position, it is basically restricted to informal conversation or casual style. Non-standard cases of omission, on the other hand, should be more frequent in spoken language, as this medium is usually considered the major locus of grammatical innovation (Kortmann 2006, 615). Therefore, a complete account of the influence of ‘Text type’ would require the inclusion of triple interactions in the regression analysis carried out in Section 5.2.3 below, such as an interaction between ‘Variety,’ ‘Text type,’ and ‘Coordination,’ but this would stretch the present dataset considerably. Since this full exploration of register and medium effects lies beyond the scope of the present study, triple interactions were not examined here. What we do expect to find with respect to text type effects, however, is a relatively high degree of variation across the three varieties considered, as stylistic effects basically boil down to social and cultural conventions (cf. Section 5.2.3.3).

Finally, individual variation is also taken into consideration. For this purpose, the instances of omitted and overt pronouns analysed were annotated for ‘Speaker,’ a variable containing a unique identifying tag for each of the individuals that contributed data to the sample. Individuals have been shown to be a major source of variation, sometimes even the most important one, in corpus studies and psycholinguistic experiments (cf. Tagliamonte and Baayen 2012). Therefore, controlling for speaker effects is of utmost importance in order to arrive at the true impact of the other constraints.

5.2.3.2 Language-internal constraints

Two language-internal factors figure prominently in the research literature on pronoun omission: agreement and context. To measure the influence of these variables, the data were annotated for the presence or absence of subject verb agreement (up to the level of English’s

morphological maximality; cf. Chapter 3, Section 3.2.1) and the accessibility of antecedents. ‘Agreement’ has three levels: present, absent, and deleted. Present and absent refer to the presence and absence of subject verb agreement morphology (cf. examples (42) and (43) respectively), while deleted refers to instances in which there should have been agreement by normative standards, but this was not overtly expressed (cf. example (44)).

(42) a. A: We’ll call **Raj** B: **Ø**_i Does not mind (ICE-IND:S1A-003#65-66)

b. **It**’s very very very quiet at the moment (ICE-GB:S1A-027#3)

(43) a. **we**_i waited for a long time for fifty-one after that **Ø**_i took the train then up till Eunós (ICE-SIN:S1A-001#137)

b. **I** speak uh Hindi (ICE-IND:S1A-029#17)

(44) a. **He**_i simply doesn’t want to do it but **Ø**_i want to show that he has done it (ICE-IND:S1A-045#76)

b. **It** sound so nice (ICE-IND:S1A-029#78)⁴³

According to Cole (2009, 2010), agreement functions as an aid to recover the morphological features of deleted pronouns and, therefore, they should be favoured when subject verb agreement morphology is present. However, Sato and Kim (2012; cf. Chapter 3, Section 3.3.2) claimed that subject omission is blocked in SgE when the verb exhibits overt agreement with the subject, thus directly contradicting the hypothesis postulated by Cole. These two positions are put to the test in the present study.

Accessibility refers to how retrievable from memory an entity or event is for the addressee in a given discourse context. The accessibility of referents was measured on the

⁴³ In these and the following examples, *a*’s provide instances of omitted pronouns and *b*’s of overt pronouns.

basis of Ariel's accessibility theory (cf. Chapter 3, Section 3.2.2), which was operationalized in the following terms. The accessibility of an entity or event increases if it is salient in the discourse and if there is a tight link between the anaphoric element and its linguistic antecedent.⁴⁴ A referent is salient, first, if it is a participant in the conversation (or the writer/reader of a text). Second, saliency increases as a function of topicality, with discourse topics being more prominent than non-topical referents. An entity was considered here to have high topicality if it was mentioned at least twice in the four clauses preceding the occurrence of the omitted/overt pronoun (cf., for instance, Toole 1996). Third, an emphasized antecedent, for instance, by being preposed, postposed, or by being part of a cleft or a pseudo-cleft sentence, is also salient. Fourth, an animate entity, especially if it is human, is more prominent for the addressee than a non-animate one and, similarly, antecedents functioning as subjects are more salient than those in other syntactic positions. Fifth, the absence of other potential antecedents in the surrounding context is also considered here to increase the saliency of the referent. Potential antecedents refer to those which are compatible with the correct referent in terms of their person, number, and animacy features (cf. Williams 1988, 351) and which are located between the anaphoric element and its correct antecedent or in the three clauses preceding the occurrence of the omitted/overt pronoun. Finally, saliency is heightened if the entity referred to was last coded by a low accessibility marker in terms of Ariel's hierarchy of accessibility-coding devices. High accessibility markers comprised overt pronouns, while low accessibility markers involved other noun phrases.

⁴⁴ Anaphoric elements that referred to the participants in the communicative exchange were considered to have exophoric reference, that is, they refer directly to an entity in the extra-linguistic context and do not have linguistic antecedents. These cases, therefore, were only analysed with respect to the saliency of their referents and not to the tightness of the link between the anaphor and its antecedent.

The link between the anaphoric element and its antecedent is tighter, first, if the anaphor and the antecedent are in embedded or coordinate clauses. Second, the two elements are more tightly linked if the sentences containing them are pragmatically cohesive, that is, if they are connected by conjunctive adverbs or if the antecedent is in an initiating speech act (e.g. a question) and the anaphoric element in a responding speech act (e.g. the answer to the question). Finally, a short distance in clauses between the anaphor and its antecedent also strengthens the connection between them. Both finite and non-finite clauses were taken into account, and utterances without a verb that were located in an independent turn were also counted, with the exception of back channel words such as *yeah* or *uhm*. There is no agreement in the literature with respect to how long an entity remains active in memory. Givón (1983) proposed an arbitrary limit of twenty clauses to the left of the anaphoric element. The present paper deals with markers of high accessibility (i.e., omitted and overt pronouns) and, therefore, only shorter distances are taken into consideration. Williams (1988) found that there is an average distance of three to four clauses between omitted pronouns and their antecedents in the speech of second language speakers of English and SgE speakers. For this reason, distances longer than four clauses are not considered here to increase the accessibility of the antecedent.

Three levels of accessibility are distinguished in the present study: high, intermediate, and low. Highly accessible referents are those that are salient and, if applicable, showed a tight link between the anaphoric element and its antecedent (cf. example (45)). The intermediate level comprised those referents which were either salient or which exhibited a tight link (cf. example (46)). Finally, referents with low accessibility were those that were not salient and did not show a tight link between the anaphor and the antecedent (cf. example (47)).

- (45) a. Then **Linda_i** rang and **Ø_i** advised me not to Fax the apology to Africa before considering the contents in case she tries to sue me for [...] defamation of character. (ICE-GB:W1B-007#45)
- b. **Two teenage girls_i** conspired to kill one of the girl's mother. **They_i** were only 15 years old. **They_i** formed a close friendship and lived half in the real world and half in fantasy. (ICE-SIN:W1B-011#57-59)
- (46) a. so I took it back to the kitchen and they **they_i** tried to convince me that this fish had swallowed the fly before it died. I said but they live under the sea you know you can't have flies under water. Ah it must have been swallowed by the uh sardine. I couldn't believe it. **Ø_i** Wouldn't give me another one. (ICE-GB:S1A-055#223-229)
- b. A: **Laura_i** likes tea bags you see after they've had taken some of the strength out. C: You don't need to explain to us. A: Bloody skinflint. B: **She_i**'s getting meaner by the day. (ICE-GB:S1A-042#44-47)⁴⁵
- (47) a. B: [...] he told us there's no place then the lady said oh if you don't mind it's directly under the air duct you know air vent so it's very cold and say you can see the weather it's cold because it was raining. She said do you mind. I said oh **it_i** doesn't matter we don't really mind because you know Fu Leong and him won't mind the cold. First they are all so big and is Alice <unclear> word </unclear> perspiring. A: Anyway you all eating hot food. [...] B: That's why Joe said **Ø_i** doesn't matter (ICE-SIN:S1A-067#157-163)
- b. Say what **he_i** told (ICE-IND:S1B-069#3)⁴⁶

⁴⁵ The referents of the omitted and overt pronouns in (47) have intermediate accessibility because they are salient (by virtue of being human subjects) but the distances between the anaphoric elements and their antecedents are longer than four clauses.

The assumption in accessibility theory is that highly accessible entities or events are easily retrieved from memory by the addressee and, therefore, they can be referred to by means of minimal – non-informative, non-rigid, and highly attenuated – forms without incurring in extra processing costs. This has a direct bearing on the hidden complexity of structures with omitted pronouns: even though the lesser explicitness of these structures has the potential of increasing their hidden complexity, this increase may be mitigated, or even completely neutralized, if the antecedents of omitted pronouns are highly accessible. Omitted pronouns, then, should be more frequently or even almost exclusively found when they are coindexed with highly accessible referents since, otherwise, the addressee would not be able to identify the entity or event being alluded to. However, since both omitted and overt pronouns are high accessibility markers and thus commonly used with this type of antecedents, referent accessibility is not expected to discriminate strongly between the two alternatives.

The next set of language-internal factors considered are related to the canonical contexts for pronoun omission in English (cf. Chapter 3, Section 3.3.3). These are ‘Coordination,’ ‘Clause position,’ and ‘Sentence type.’ ‘Coordination’ has two levels: coordination, that is, when the omitted/overt subject pronoun occurs in the second or successive conjuncts of coordinate clauses linked by the conjunctions *and*, *or*, and *but* and it is coreferential with the subject of the first conjunct (cf. example (48)), and no coordination (cf. example (49)).

- (48) a. [...] so **he_i** threw a piece of brick you know and **Ø_i** made him stand or made him wait there [...] (ICE-IND:S1A-004#18)

⁴⁶ In (47a), several potential antecedents are found between the omitted pronoun and the correct antecedent, so the referent is not salient. In addition, they are separated by more than four clauses. In (47b), the referent was not mentioned before so it is not highly accessible.

b. **Hei**'s got a new one and **he**_i's quite happy with it (ICE-SIN:S1A-095#207)

(49) a. In a way I wish **he had come round yesterday**_i. [...] **Ø**_i Would have been sweet
(ICE-GB:S1A-042#28-30)

b. So when whenever I ask him **he**_i may not do it promptly [...] **he**_i said do it later
(ICE-SIN:S1A-015#46-48)

Since, according to reference grammars, omission is very common in coordinate contexts, this situation is expected to hold in the present study.

‘Clause position’ involves the location of the omitted/overt pronoun in the clause, which may be initial, as in (50), or non-initial, as illustrated in example (51):

(50) a. **Ø** Have to close my eyes for a stop... (ICE-GB:W1B-010#114)

b. **I** have already communicated you that you have to accept 1000/- Rs. for the proposal. (ICE-IND:W1B-003#11)

(51) a. Haan onion **we**_i [...] make it red. Then **Ø**_i put some haldi powder and some little bit of chilli as per the taste (ICE-IND:S1A-007#94-95)

b. [...] we've been uh talking to various groups of people and uh trying to see whether there are any major issues that and unhappiness **they**_i have uh. So far **they**_i have brought out very few (ICE-SIN:S1B-044#71-72)

The choice of operationalizing position in terms of clauses, rather than, for instance, turns or intonation units, is due to the inclusion in the present study of both spoken and written language. Initial position, especially in casual conversation, was another one of the contexts mentioned in reference grammars in which omission is allowed in English. Therefore, a

similar distribution is expected here, with omitted pronouns being preferred in initial clause position.

‘Sentence type’ refers to whether the sentence in which the omitted/overt pronoun occurs is declarative, as exemplified in (52), or interrogative, as in (53):

(52) a. **He_i** speaks 3 languages and **Ø_i** lives all over Europe (ICE-GB:W1B-003#187)

b. **I** will call up from China or Japan next time sometime in the middle of May (ICE-IND:W1B-010#11)

(53) a. Uhn when are **Ø** going to get married Mr Shellroy? (ICE-IND:S1A-024#106)

b. So have **you** actually started looking around or not for your gown and everything else (ICE-SIN:S1A-001#9)

Torres Cacoullos and Travis (2014) found that 1st person singular pronouns are categorically overtly expressed in interrogative sentences in AmE conversation. This is clearly not the case in the present dataset, where several instances of omitted pronouns occurred in questions (cf. Section 5.3 below), but omission is still expected to be favoured in declarative sentences.

Priming or persistence effects have also been consistently found in studies of pronoun omission. ‘Priming,’ operationalized here as the form of the previous coreferential subject of the relevant omitted/overt pronoun in the preceding five clauses, is a categorical variable with three levels: omitted pronoun (cf. example (54)), overt pronoun (cf. example (55)), and noun phrase/no priming (cf. example (56)).

(54) a. I’ve actually called the Indian High Commission in S’pore (**Ø_i** says **Ø_i** can’t help)

[...] (ICE-SIN:W1B-007#8)

b. Ø_i Can't remember. Anyway I_i'm actually looking forward to this evening at home so that I can pray! (ICE-GB:W1B-003#80-81)

(55) a. but he_i says it in this really furtive little voice. Ø_i Doesn't want anyone else to hear. (ICE-GB:S1A-055#100-101)

b. At night, while drinking the milk, he_i found instead a trickle of sour curd. Tearing open the bag, he_i found the milk had coagulated. (ICE-IND:W2D-018#76-77)

(56) a. After recording the different ingredients [...], **the Sound Engineer**_i puts them all together in the required sequence and proportion, in other words, Ø_i edits the programme on the directions of the Producer. (ICE-IND:W2B-014#101)

b. I'll have to stop **talking about the place**_i, it_i's bringing tears to my cheeks. (ICE-GB:W1B-001#79-80)

If the previous coreferential subject is an omitted pronoun, then omission should be favoured. This is not expected to be the case when the omitted pronoun is not primed, that is, when the last coreferential subject is separated from the omitted pronoun by more than five clauses or when it is realized as an overt pronoun or any other noun phrase.

The type of clause in which the omitted/overt pronoun occurs, that is, main (cf. example (57)) versus subordinate (cf. example (58)), is also expected to influence the choice between the two alternatives.

(57) a. More importantly, Ø_i hope I do not become your stumbling block. (ICE-SIN:W1B-001#82)

b. On a more humorous note, how are your studies (dissertation) coming along, nicely I trust. (ICE-GB:W1B-001#21-22)

- (58) a. [...] we can first check out the activities [...] then on the basis of activities **we**_i can calculate how many students **Ø**_i will uh actually require (ICE-IND:S1B-075#52)
- b. So I tape the ending of Friday Background **the music**_i. I'm not sure if **it**_i's the same as the opening music (ICE-SIN:S1A-015#135-136)

Rohdenburg's (1996, 151) complexity (or transparency) principle states that "[i]n the case of more or less explicit grammatical options the more explicit one(s) will tend to be favored in cognitively more complex environments" (cf. also Mondorf 2009). Subordinate clauses have been shown to involve an increased processing load (e.g. Hawkins 1999; and references therein) and, therefore, can be characterized as cognitively complex environments. Omitted pronouns, being a clearly less transparent grammatical option than overt pronouns, should be dispreferred in subordinate clauses. Torres Cacoullos and Travis (2014) excluded subordinate clauses from their data since they found no instances of omission in these contexts in AmE conversation. This is not the case in my dataset, and so both main and subordinate clauses were considered.

The frequency of pronoun omission also depends on the actual pronominal form that is deleted. To capture the potential variability associated with the different personal pronouns, the data were also annotated for a variable 'Pronoun' with the following levels: speaker/hearer, including 1st and 2nd person pronouns, and other (3rd person pronouns), illustrated in examples (59) and (60):

- (59) a. **Ø** Hope it will not last long. (ICE-IND:W1B-011#88)
- b. **I** hope everything has turned out well for you both [...] (ICE-GB:W1B-008#145)

- (60) a. Oh, be4 I forget, “**Chitrai**” sends you her love. **Øi** Has been asking about you since you left. (ICE-SIN:W1B-003#4-5)
- b. **Shei** has taught at school and college level English literature (ICE-IND:S1B-048#5)

Pronouns referring to speakers or hearers were shown by Claes (2017; cf. Chapter 3, Section 3.2.3) to be coded by overt pronouns more frequently than those referring to other entities in Cuban Spanish. He hypothesized that this distribution is connected to the fact that the interlocutors in a conversation are cognitively more prominent for speakers, and more prominent entities favour overt coding. If this is true, we should find a similar distribution in the present study, with 1st and 2nd person pronouns opting for overt expression more frequently than 3rd person pronouns.

As seen in Chapter 3, Section 3.2.3, frequency of co-occurrence between the subject pronoun and the following verb might have an effect on the probability of omission. Speakers seem to rely on their past linguistic experience and favour overt expression when a pronoun and a verb co-occur very frequently. Two numerical variables are included in the present study so as to assess the influence of frequency effects: ‘Co-occurrence frequency,’ which simply reflects the number of occurrences of a specific subject pronoun followed by a particular verb form, and ‘Collocational strength,’ which measures how strongly associated the pronoun and the verb are. To avoid circularity, these measures were not extracted from the same corpus from which the data were retrieved but from the *Corpus of Global Web-Based English* (Davies 2013), or *GloWbE*, a much larger database containing almost two billion words from twenty varieties of English. ‘Co-occurrence frequency’ basically indicates the number of times (normalized per million words) a pronoun and a verb co-occur in the British,

Indian, and Singaporean components of *GloWbE*: *I am*, for instance, has a normalized frequency of 546.23 in BrE, 583.13 in IndE, and 613.48 in SgE. ‘Collocational strength’ was computed using Gries’ (2007) *R* script, which calculates the strength of association between two or more words (or between words and constructions) on the basis of contingency tables such as Table 10. The output of the script provides several association measures reflecting the collocational strength between, in this case, pronouns and verb forms. The results reported here are based on calculating the odds ratio of each pronoun-verb combination. For instance, the odds ratio of *I am* in IndE is 2.42, which derives from dividing the odds of finding *I* given *am* (i.e., 56,232 divided by 34,802) by the odds of finding *I* given any other word (i.e., 590,029 divided by 95,749,825) and transforming the resulting value (262.15) into the logarithmic scale. The higher this value is, the stronger the association of the relevant pronoun-verb combination, which leads to a higher chance of being stored in memory as a complex unit and thus a lower probability of omission. Pronouns and verbs with lower co-occurrence frequency and collocational strength values, on the other hand, should present a higher incidence of subject omission.

Table 10. Contingency table for *I am* in IndE

	<i>I</i>	Other	Total
<i>Am</i>	56,232 ^a	34,802 ^b	91,034 ^c
Other	590,029 ^d	95,749,825 ^e	96,339,854 ^f
Total	646,261 ^g	95,784,627 ^h	96,430,888 ⁱ

^a Frequency of *I am*.

^b Frequency of *am* preceded by any subject but *I*.

^c Total frequency of *am*.

^d Frequency of *I* followed by any verb form but *am*.

^e Number of words in the corpus minus *I* and *am*.

^f Number of words in the corpus minus *am*.

^g Total frequency of *I*.

^h Number of words in the corpus minus *I*.

ⁱ Total number of words in the corpus.

Frequency effects are also explored by means of a categorical variable coding the class of the verb following the omitted/overt subject pronoun. Three levels are distinguished: lexical verbs, as in (61), non-modal auxiliaries, illustrated in example (62), and modal auxiliaries (cf. example (63)):

- (61) a. **Dr. Karande_i** rang me up a week before and **Ø_i** informed me of your comfortable Pune situations. (ICE-IND:W1B-003#79)
- b. So **we** have one hundred thousand pages okay (ICE-SIN:S1B-074#9)
- (62) a. **Ø** Am having a 2-day break before assuming invigilation for A Maths. (ICE-SIN:W1B-003#182)
- b. Oh dear, **I**'m rather late but Happy Birthday, anyway. (ICE-GB:W1B-010#125)
- (63) a. A: And **the people_i** will be evacuated [...] B: **Ø_i** Will be alerted (ICE-IND:S1A-029#118-121)
- b. **I**'d like to answer that in a s in a slightly different way (ICE-GB:S1A-001#118)

Non-modal auxiliaries and personal pronouns co-occur very frequently: in the present sample, they exhibit a mean normalized co-occurrence frequency of 778.77, which means that non-modal auxiliaries are preceded by a subject pronoun an average of 778.77 times (per million words) in *GloWbE*. This value is much higher than those showed by modal auxiliaries (191.24) or lexical verbs (56.77). The high co-occurrence frequency of pronouns and non-modal auxiliaries may have to do with general principles of communication and semantics (Bybee 2007, 18). First, people usually talk about themselves and the human beings around them, and personal pronouns are the device most commonly used to refer to this type of antecedents. In addition, elements with abstract and general meanings, such as auxiliary

verbs, can occur in combination with many different words. Therefore, as suggested by Bybee (2001, 356), sequences of words made up by a pronoun plus a form of an auxiliary may be stored in memory as single units due to their high frequency of co-occurrence. This analysis is supported by the fact that non-modal auxiliaries often appear in a contracted form when they are preceded by personal pronouns (cf. Krug 1998). These units formed by a pronoun plus a non-modal auxiliary are accessed and retrieved from memory as a whole, that is, the pronoun and the verb are no longer independent items that can be individually selected and omitted. As a result, pronoun omission is predicted to be disfavoured when followed by a non-modal auxiliary. Lexical verbs, on the other hand, should not inhibit omission, and modal verbs should occupy an intermediate position.

Finally, each data point in the set was annotated for the specific verb form with which it occurs and its corresponding lemma so as to control for lexical effects. As in the case of individual speaker effects, discussed in Section 5.2.2.1, lexical items have been shown to be a major source of variation (cf. Gries 2016). Therefore, ‘Verb form’ and ‘Verb lemma’ were included as random effects in the mixed-effects logistic regression that was carried out to assess the influence of the variables described in the present section on the choice between omitted and overt referential subject pronouns.

5.2.3.3 Interim summary

Table 11 summarizes the variables included in the present analysis, as well as the corresponding hypotheses postulated for each of them (‘>’ stands for ‘higher frequency of

omission than’).⁴⁷ In brief, referential subject pronoun omission is predicted to increase in frequency in the high-contact varieties and to be preferred in omission-friendly environments, that is, when independent processing/cognitive factors favour deletion. In addition, it is also hypothesized to be more likely in the same contexts in which it commonly occurs in standard varieties, such as coordination, initial position, and declarative sentences. The hypothesized spread of omission in SgE and IndE, however, is not expected to occur to the same extent in all the conditions included in Table 11. Constraints firmly grounded in processing considerations have been shown to be rather stable across varieties since all users have the same cognitive architecture and are thus subject to similar limitations in online production (Szmrecsanyi et al. 2016; Heller, Szmrecsanyi, and Grafmiller 2017; Röthlisberger, Grafmiller, and Szmrecsanyi 2017). Register and medium of production, on the other hand, tend to show more variation, as stylistic differences are essentially a matter of social and cultural conventions.

In line with previous research, we expect that the speakers of the three varieties examined here behave similarly with respect to the influence of ‘Agreement,’ ‘Accessibility,’ and ‘Sentence type,’ ‘Priming,’ ‘Clause type,’ ‘Pronoun,’ ‘Co-occurrence frequency,’ ‘Collocational strength,’ and ‘Verb class,’ all clearly cognitive- or processing-based constraints. ‘Text type,’ ‘Coordination,’ and ‘Clause position,’ on the other hand, are hypothesized to exhibit variation across the three varieties, with SgE and IndE favouring omission vis-à-vis BrE in some of these contexts.

⁴⁷ ‘Speaker,’ ‘Verb form,’ and ‘Verb lemma’ are not included in Table 11 because no hypotheses were put forward for these variables. They are considered in the analysis to control for individual speaker and lexical item preferences that are not accounted for by any of the other variables.

Table 11. Variables included in the analysis

Variable		Hypothesis
<i>External</i>	Variety	SgE/IndE > BrE
	Text type	Variation across the three varieties expected
<i>Internal</i>	Agreement	Present > Absent/Deleted (Cole 2009, 2010); Absent/Deleted > Present (Sato and Kim 2012)
	Accessibility	High > Intermediate > Low
	Coordination	Coordination > No coordination
	Clause position	Initial > Non-initial
	Sentence type	Declarative > Interrogative
	Priming	Omitted > Overt/NP & No priming
	Clause type	Main > Subordinate
	Pronoun	Speaker/hearer > Other
	Co-occurrence frequency	The higher, the fewer instances of omission
	Collocational strength	The higher, the fewer instances of omission
	Verb class	Lexical > Modal aux. > Non-modal aux.

5.2.4 Statistical analysis

Both univariate and multivariate analyses of the data are conducted. The variable of interest, in both cases, is ‘Omission,’ that is, whether the referential subject pronoun is realized in omitted or overt form. First, we inspect the distribution of these two alternatives in each of the contexts mentioned in the previous section separately, that is, how frequently each of them occurs in, for instance, main versus subordinate clauses or declarative versus interrogative sentences. The results are visualized in frequency tables and plots.

Second, multivariate analyses are carried out to assess the independent effect of each of the variables on the probability of encountering an omitted pronoun while simultaneously controlling for the effects of the rest. To this purpose, the data are first analysed by means of a binary mixed-effects logistic regression using the `lme4` package (Bates et al. 2015) so as to discover which factors have a statistically significant influence on the dependent variable ‘Omission,’ as well as the direction and strength of that influence. Afterwards, per variety random forests are conducted with the `cforest()` function in the `party` package (Hothorn et al. 2006; Strobl et al. 2007, 2008) to measure the relative importance of the significant factors in each of the varieties’ pronoun-deletion grammar. These two procedures, which are explained in Sections 5.2.4.1 and 5.2.4.2 respectively, are used in the present study to quantify the structural and system complexities of the varieties at hand.

5.2.4.1 Regression modelling and structural complexity

Binary logistic regression is a statistical test that aims at estimating the likelihood of the levels of a binary dependent variable, such as the occurrence of omitted versus overt referential subject pronouns, on the basis of a series of independent variables or predictors (Baayen 2008, 195; Levshina 2015, 253). Mixed-effects models include random as well as fixed effects (Baayen 2008, 241–42; Tagliamonte and Baayen 2012). Predictors with repeatable levels, such as ‘Clause type,’ are fixed: if we drew a sample of omitted and overt subject pronouns from a different corpus, they would still be classifiable in terms of whether they occurred in a main or in a subordinate clause. ‘Speaker,’ on the other hand, is a random predictor since, if we retrieved new data from a different corpus, it would inevitably contain a new sample of speakers with different individual preferences for omitted or overt pronouns. Mixed-effects models accommodate the inclusion of, for instance, speakers and lexical items

as random effects, thus allowing each speaker and lexical item to have a different baseline preference for, in this case, the omitted or overt referential subject pronoun variant.

In the present study, all the predictors in Table 11 above are included in the model as fixed effects, while ‘Speaker,’ ‘Verb form,’ and ‘Verb lemma’ are included as random effects. This setting of predictors enables us to examine in which contexts pronoun omission is (dis)favoured while at the same time controlling for speaker and lexical effects. In addition, interactions of ‘Variety’ with the other fixed predictors are also considered to discover potential differences between BrE, IndE, and SgE as to the influence of the language-internal and -external constraints on the choice between omitted and overt pronouns. This is a complex model which results in a total number of 42 regression coefficients. As a rule of thumb, the number of observations in the dataset for the minority variant should be at least ten to fifteen times larger than the number of coefficients (Baayen 2008, 195; Levshina 2015, 143–44). The present data comprise 1,229 tokens of omitted pronouns and the same number of overt pronouns, which allows the inclusion of 82 to 123 coefficients in the model. The final model, after discarding non-significant predictors and interactions, contains 26 regressor coefficients, which means that we are well within the limits of even the strictest of rules (Baayen 2008, 221).

The process for the selection of the final model follows the steps in Heller, Szmrecsanyi, and Grafmiller (2017; cf. also Zuur et al. 2009; Gries 2015). A full model with all fixed/random predictors and interactions is first considered. After that, non-significant predictors are gradually removed, first random effects and then fixed effects. Finally, the model is validated using the bootstrapping method described in Baayen (2008, 282–84), whereby the same regression formula is applied to a hundred random samples of the data.

This procedure results in a range of values which, with a 95% probability, contains the real value of each coefficient.

Finally, the partial effects of the predictors, that is, the effect of each predictor when all the other effects in the model are held constant, are visualized using the `ggplot2` package (Wickham 2009). The results of the binary mixed-effects logistic regression are interpreted here in terms of structural complexity. A preference for omitted subject pronouns translates into a decrease in the structural complexity of a variety. Similarly, if the omitted variant is favoured in a particular context in all three varieties but this preference is stronger in one variety as compared to the others, this variety is considered to be structurally simpler than the rest. Exceptions to this interpretation are those contexts mentioned in Section 5.2.2 in which omission results in a heightened processing load. A preference for omitted pronouns in these environments is considered to increase structural complexity.

5.2.4.2 Random forests and system complexity

As in the case of regression modelling, the ultimate goal of random forests is to estimate the probability of the levels of a dependent variable given a series of predictors. Random forests are constructed on the basis of many conditional inference trees, which make a series of binary splits in the data (Tagliamonte and Baayen 2012, 159; Levshina 2015, 292). First, the predictor that is more strongly correlated with the dependent variable is selected and, on the basis of this predictor, the algorithm examines whether dividing the data in two creates two new subsets of cases in which one of the levels of the dependent variable is more probable than the other. This process is then repeated for each new partition until no further significant splits are found. Each tree in the forest is computed from a random sample of datapoints and variables. To assess the relative importance of the predictors included in the model, the

algorithm disrupts any potential association between the dependent variable and a given predictor by randomly permutating its values (Tagliamonte and Baayen 2012, 160). Table 12 illustrates this procedure with a hypothetical distribution of ‘Omission’ and ‘Clause type’:

Table 12. Example of permutation in random forests

Omission	Clause type (observed)	Clause type (permuted)
Omitted	Main	Subordinate
Overt	Subordinate	Subordinate
Overt	Subordinate	Main
Omitted	Main	Main
Omitted	Main	Subordinate
Overt	Subordinate	Main
Overt	Subordinate	Subordinate
Omitted	Main	Main

In this constructed example, there is a clear association between omitted pronouns and main clauses. However, after randomly changing the values of the predictor (cf. the column ‘Clause type (permuted)’), this association is broken: omitted pronouns no longer occur in main clauses in each and every instance if we consider the permuted version of ‘Clause type’. The relative importance of a given predictor is then calculated as the degree to which the model becomes worse when computed with the permuted version instead of the original predictor. In the constructed example of Table 12, the model would become substantially worse with the permuted version of ‘Clause type,’ as this predictor would not accurately discriminate between omitted and overt pronouns anymore. Finally, a random forest provides an ordering of variables as to their relative importance in the model, which is a measure of how well each

predictor discriminates between the levels of the dependent variable. This final ordering of predictors is the one that is supported by the largest number of conditional inference trees.

In the present study, per variety random forests are computed with all the predictors that are found to be statistically significant by the binary mixed-effects logistic regression. This allows us to assess which constraints actually play an important role in each of the varieties' pronoun-deletion grammars and which ones are more superfluous, that is, which predictors do not have any discriminatory power between the omitted and overt pronoun variants. As a rule of thumb, only those variables with an importance value that differs from the absolute value of the variable with the smallest score are considered to be relevant.

In order to conduct the random forest analyses, the dataset was divided into three subsets, each one containing data from one of the varieties. Each of these subsets includes an equal number of instances of omitted and overt pronouns, amounting to 728 observations in BrE, 650 in IndE, and 1,080 in SgE.⁴⁸

The results of the per variety random forests provide an indication of the varieties' system complexity. The fewer the constraints in a variety that play a role in the choice between omitted and overt pronouns, the simpler its pronoun-deletion grammar is and the easier it is to learn. On the other hand, if a large number of constraints have a high variable importance value in a given variety, then it exhibits a more complex system as this requires speakers to learn a larger number of rules in order to be able to correctly use the set of referential expressions available in the grammar of their variety. This view of system complexity is inspired by Shin (2014, 305), who argues that “[t]he loss of a linguistic factor that constrains linguistic choice is a type of simplification, while the emergence of a new

⁴⁸ In the case of the SgE subset, it was necessary to retrieve and annotate 134 additional instances of overt pronouns from *ICE-SIN*, as the full dataset contained 540 observations of omitted pronouns but only 406 of overt pronouns.

factor is a type of complexification”. Shin’s claim is based on Dahl’s (2004, 21) method of measuring linguistic complexity as the description length of a phenomenon (cf. Chapter 2, Section 2.3.2): a linguistic phenomenon in which many constraints are necessary to account for variation between different grammatical options requires a longer description and is thus more complex (cf. also Davydova 2010, 103; Huber 2012).

5.3 FREQUENCY OF OMITTED AND OVERT PRONOUNS PER VARIETY

This section deals with the distribution of omitted and overt referential subject pronouns in each of the contexts described in Table 11 above and in each of the three varieties examined. The results are first described in Section 5.2.1 and then discussed in Section 5.3.2.

5.3.1 Results

Starting with the language-external variables, Table 13 and Figure 14 display the distribution of omitted and overt pronouns in each of the three varieties considered, which is statistically significant according to a chi-squared test of independence ($\chi^2 = 31.04$, d.f. = 2, $p < 0.001$). SgE is the variety with the highest percentage of omitted pronouns (57.08%), followed by IndE (46.17%) and then BrE (45.05%). BrE and IndE behave very similarly and cannot be said to differ in this respect, since they exhibit almost the same proportion of omitted pronouns. In addition, in both varieties overt pronouns are the most frequent alternative. So far, it seems that, in terms of structural complexity, IndE patterns more closely with BrE, the native low-contact variety, than with SgE. However, as will be made evident in Section 5.4.1, there are actually subtle but important differences between them regarding the effects of some of the constraints that the type of analysis carried out so far cannot capture.

Table 13. Distribution of omitted and overt pronouns per level of ‘Variety’

Variety	Omitted	Overt
BrE	45.05% (364)	54.95% (444)
IndE	46.17% (325)	53.83% (379)
SgE	57.08% (540)	42.92% (406)

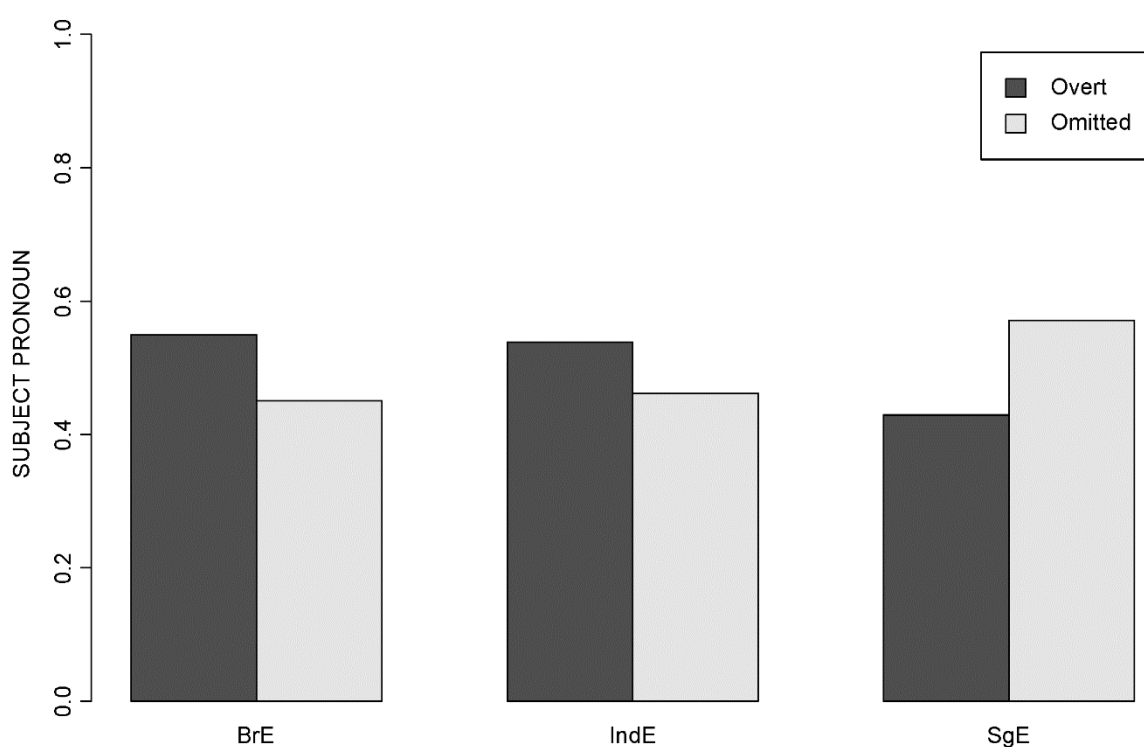


Figure 14. Proportion of omitted and overt pronouns per level of ‘Variety’

The distribution of omitted and overt pronouns in each of the text types distinguished in the present study is shown in Table 14 and Figure 15. Surprisingly, omitted pronouns occur more frequently in written formal texts (74.31% in BrE, 75.56% in SgE, and 78.85% in SgE), followed, particularly in SgE, by written informal texts (53.82% in BrE, 55.90% in IndE, and 70.39% in SgE). In spoken language, omitted pronouns are less frequent in all three varieties

and, in BrE and SgE, more so in spoken informal (31.50% and 42.26%, respectively) than in spoken formal texts (30.57% and 38.10%, respectively). Contrarywise, in IndE omitted pronouns are more frequent in spoken formal (37.16%) than in spoken informal texts (33.90%). Table 14 and Figure 15 suggest a division in terms of medium of production rather than formality: omitted pronouns are more common in writing than in speech.

As regards the distribution of omitted and overt pronouns with respect to the language-internal constraints, the distribution of omitted and overt pronouns in the three varieties already enables us to reject the two hypotheses postulated in previous research in connection with the effect of ‘Agreement.’ As shown in Table 15 and Figure 16, the presence of subject verb agreement does not favour omission. In fact, with a percentage of 55.13% in BrE, 54.33% in IndE, and 57.76% in SgE, overt pronouns are more frequent than omitted pronouns when the verb explicitly agrees with the subject. Omission is more common when there is no subject verb agreement (44.84% in BrE, 45.98% in IndE, and 59.87% in SgE) and, especially, when there should have been agreement by normative standards but it is not overtly realized. However, these results must be taken with care due to the low number of instances occurring with deleted agreement markers and, in fact, as shown by the multivariate analysis of the data (cf. Section 5.4.1.1), agreement does not have a significant effect in any of the varieties.

Table 14. Distribution of omitted and overt pronouns per level of ‘Text type’

Text type	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Spoken informal	31.50% (80)	68.50% (174)	33.90% (80)	66.10% (156)	42.26% (131)	57.74% (179)
Spoken formal	30.57% (48)	69.43% (109)	37.16% (68)	62.84% (115)	38.10% (56)	61.90% (91)
Written informal	53.82% (155)	46.18% (133)	55.90% (109)	44.10% (86)	70.39% (271)	29.61% (114)
Written formal	74.31% (81)	25.69% (28)	75.56% (68)	24.44% (22)	78.85% (82)	21.15% (22)

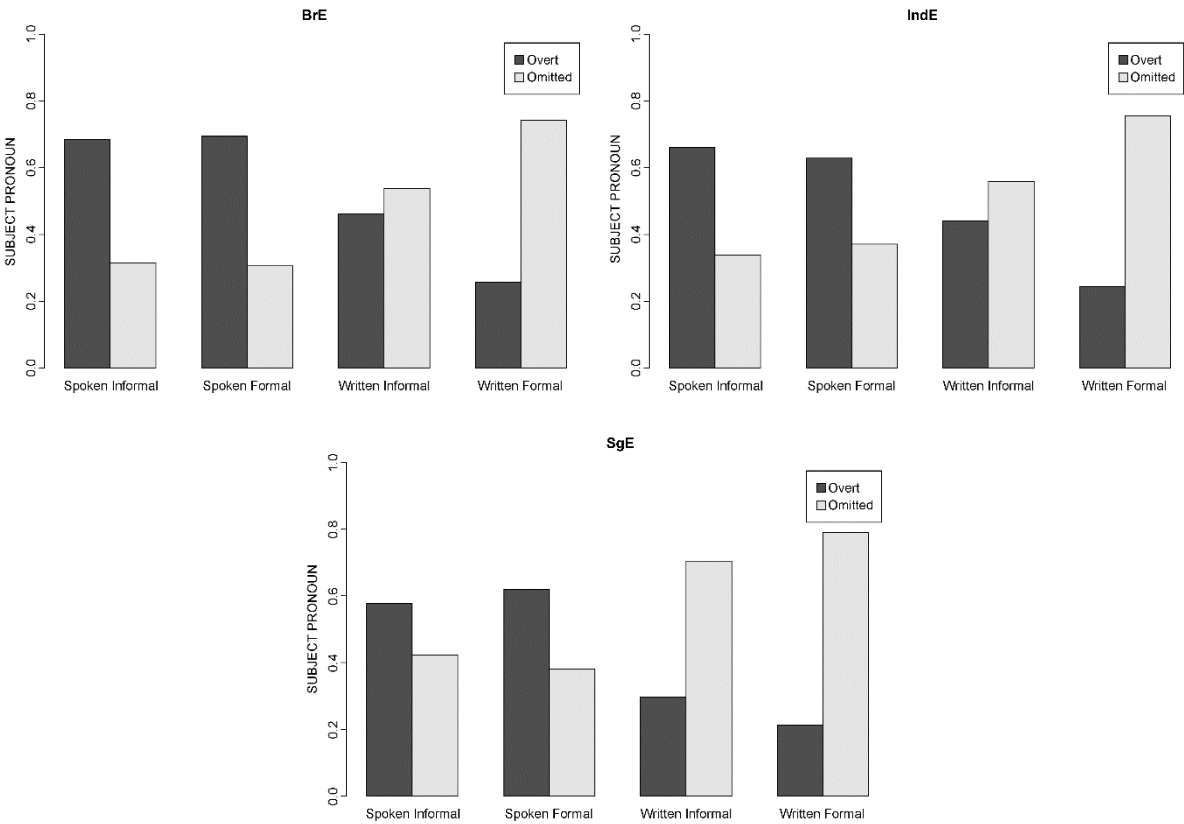
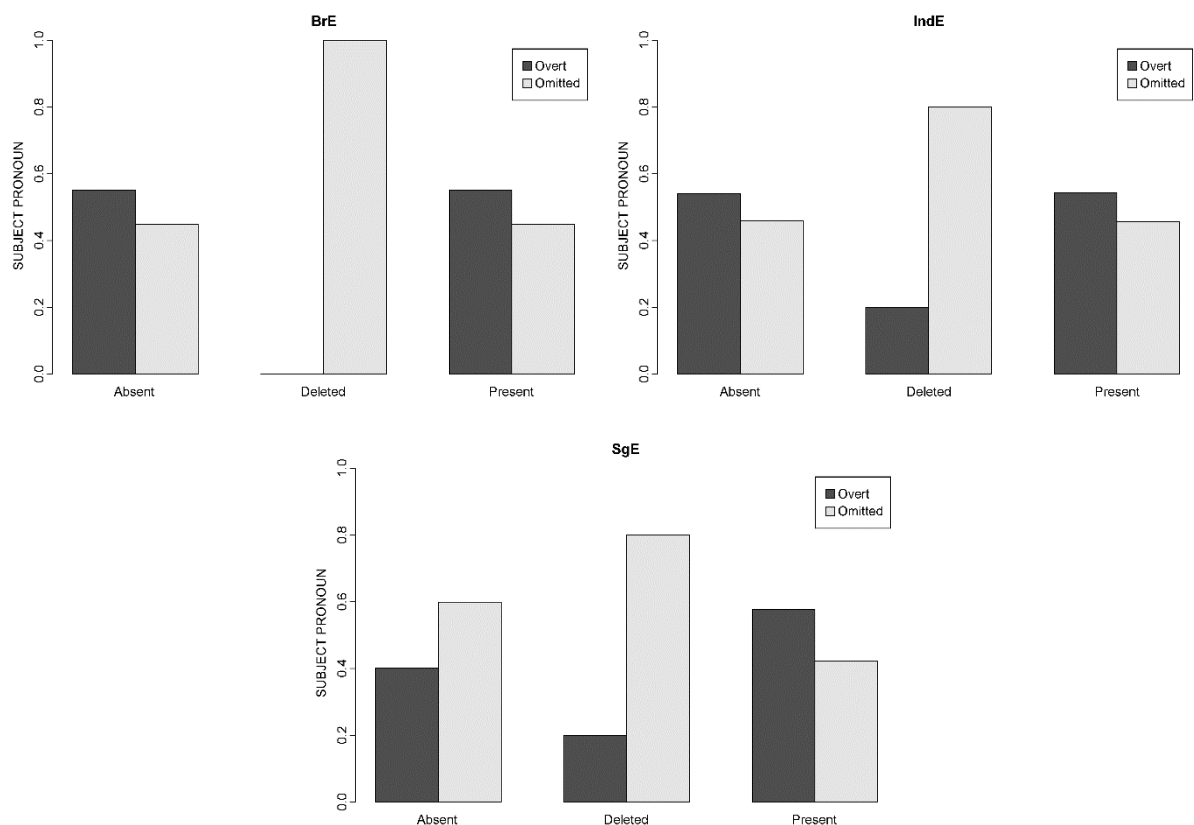


Figure 15. Proportion of omitted and overt pronouns per level of ‘Text type’

Table 15. Distribution of omitted and overt pronouns per level of ‘Agreement’

Agreement	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Absent	44.84% (291)	55.16% (358)	45.98% (263)	54.02% (309)	59.87% (464)	40.13% (311)
Deleted	100.00% (3)	0.00% (0)	80.00% (4)	20.00% (1)	80.00% (8)	20.00% (2)
Present	44.87% (70)	55.13% (86)	45.67% (58)	54.33% (69)	42.24% (68)	57.76% (93)

**Figure 16. Proportion of omitted and overt pronouns per level of ‘Agreement’**

With regard to the effect of ‘Accessibility,’ Table 16 and Figure 17 display the frequency of omitted and overt pronouns in terms of the accessibility of their antecedents. As expected, omitted pronouns are clearly disfavoured when they refer to antecedents with intermediate (30.30% in BrE, 16.98% in IndE, and 36.36% in SgE) and, especially, low (0% in BrE and IndE, and 31.82% in SgE) accessibility, that is, when their referents are judged by the speaker not to be easily retrievable. Highly accessible antecedents, on the other hand, favour omission in the two high-contact varieties and occur very frequently in BrE (47.71% in BrE, 50.32% in IndE, and 59.03% in SgE), although overt pronouns are also commonly used to refer to this type of referents (52.29% in BrE, 49.68% in IndE, and 40.97% in SgE).

Table 16. Distribution of omitted and overt pronouns per level of ‘Accessibility’

Accessibility	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
High	47.71% (344)	52.29% (377)	50.32% (316)	49.68% (312)	59.03% (513)	40.97% (356)
Intermediate	30.30% (20)	69.70% (46)	16.98% (9)	83.02% (44)	36.36% (20)	63.64% (35)
Low	0.00% (0)	100.00% (21)	0.00% (0)	100.00% (23)	31.82% (7)	68.18% (15)

5. Complexity variance in English: Pronoun omission and language contact

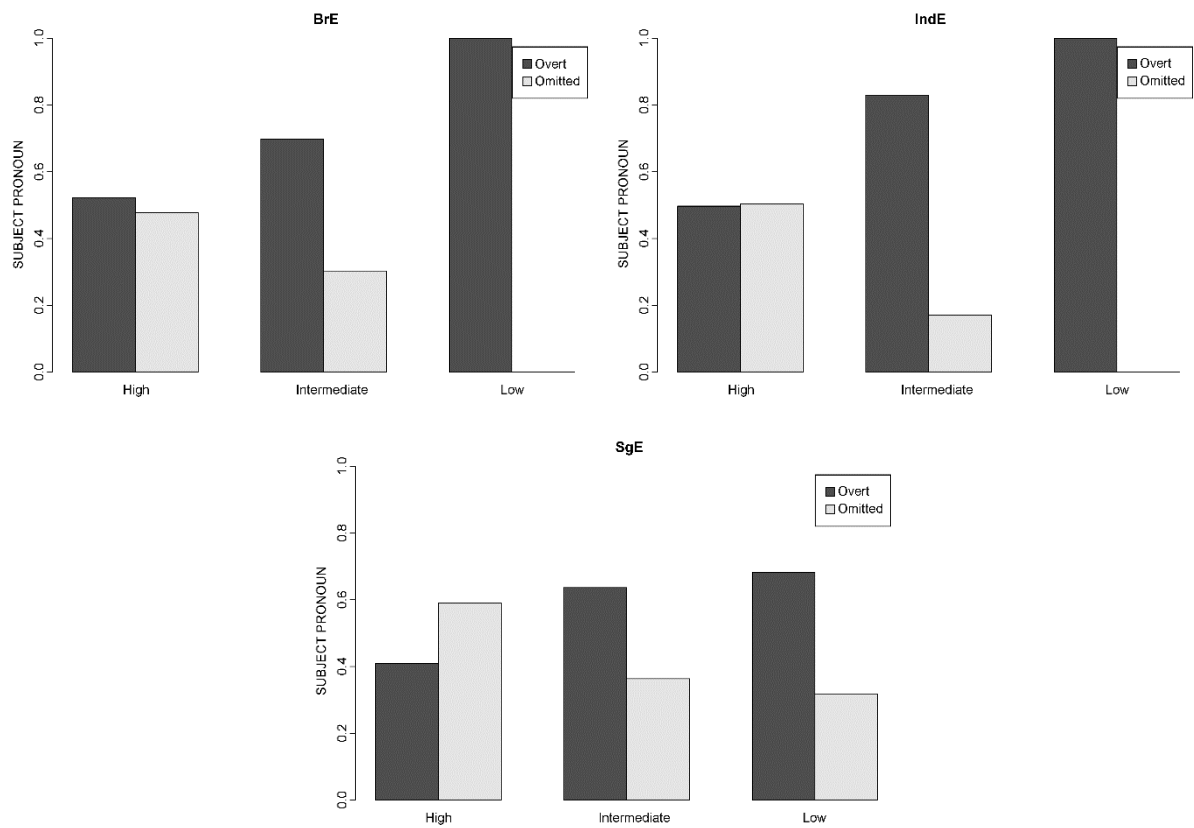


Figure 17. Proportion of omitted and overt pronouns per level of ‘Accessibility’

Interesting distributions are also found for the variables ‘Coordination,’ ‘Clause position,’ and ‘Sentence type,’ that is, those including the contexts in which omitted pronouns tend to occur in English according to reference grammars. As shown in Table 17 and Figure 18, coordinate clauses in which the subject of the second or subsequent conjuncts are coreferential with that of the first one are the omission contexts par excellence: in almost all the coordinate clauses analysed here the subject pronoun is realized in omitted form (94.14% in BrE, 94.92% in IndE, and 95.48% in SgE). In non-coordinate clauses, on the other hand, omission is less frequent than the overt expression of the subject pronoun (20% in BrE, 21.58% in IndE, and 42.31% in SgE). In absolute terms, however, there is a substantial

number of omitted pronouns in non-coordinate contexts, particularly in SgE, which means that ‘Coordination’ is not the only factor that matters.

Table 17. Distribution of omitted and overt pronouns per level of ‘Coordination’

Coordination	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Coordination	94.14% (257)	5.86% (16)	94.92% (224)	5.08% (12)	95.48% (251)	4.56% (12)
No coordination	20.00% (107)	80.00% (428)	21.58% (101)	78.42% (367)	42.31% (289)	57.69% (394)

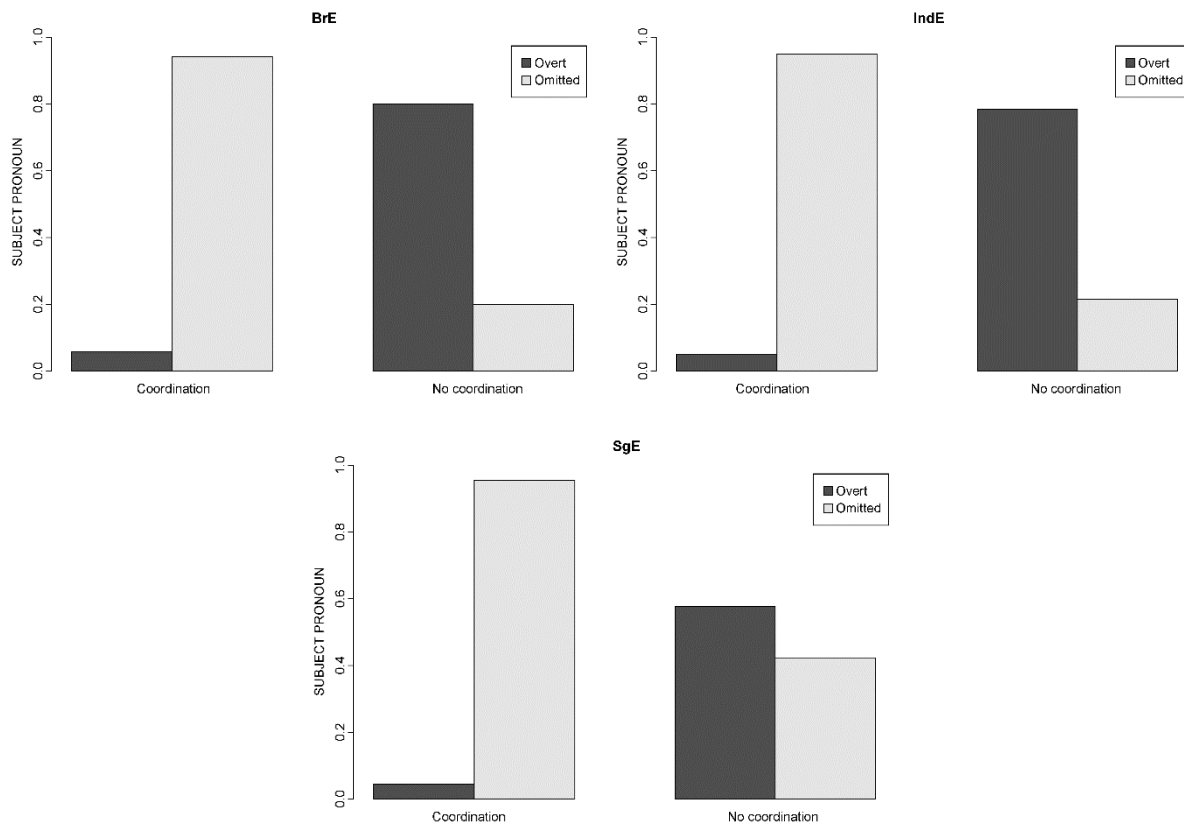


Figure 18. Proportion of omitted and overt pronouns per level of ‘Coordination’

With respect to ‘Clause position,’ omission is again more frequent in the canonical context. As can be seen in Table 18 and Figure 19, omitted pronouns occur in initial clause position with a frequency of 62.82% in BrE, 64.27% in IndE, and 70.18% in SgE, while only in the range of 8.43% to 26.24% of the cases in non-initial position. Initial position is also the context with the largest number of instances of omission in absolute terms (342 in BrE, 286 in IndE, and 466 in SgE), that is, deleted pronouns are clearly favoured in this environment.

Table 18. Distribution of omitted and overt pronouns per level of ‘Clause position’

Clause position	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Initial	62.52% (342)	37.48% (205)	64.27% (286)	35.73% (159)	70.18% (466)	29.82% (198)
Non-initial	8.43% (22)	91.57% (239)	15.06% (39)	84.94% (220)	26.24% (74)	73.76% (208)

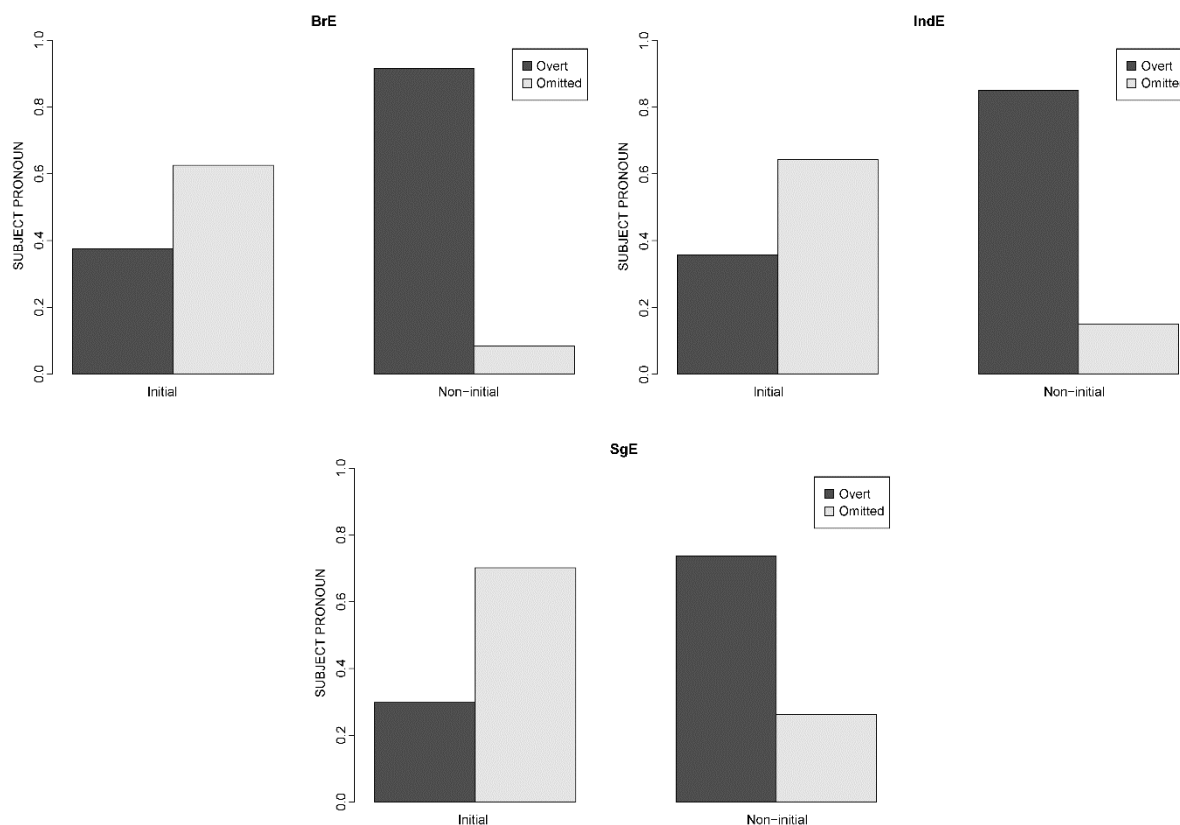
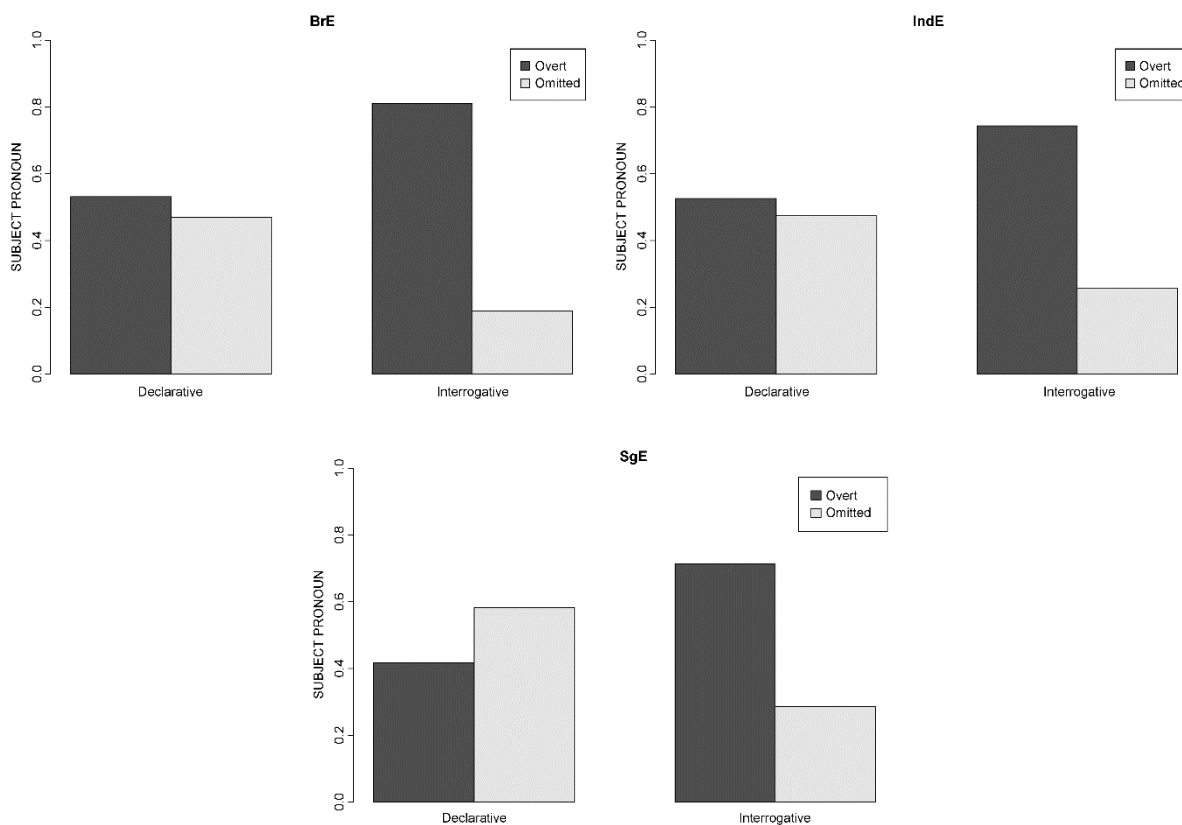


Figure 19. Proportion of omitted and overt pronouns per level of ‘Clause position’

Similarly, as displayed by Table 19 and Figure 20, omitted pronouns are more common in declarative (46.89% in BrE, 47.37% in IndE, and 58.18% in SgE) than in interrogative sentences (18.87% in BrE, 25.64% in IndE, and 28.57% in SgE), that is, omission is disfavoured in non-declarative contexts. However, overt pronouns are also more frequent in absolute terms in declarative than in interrogative sentences, as most instances (401 in BrE, 350 in IndE, and 381 in SgE) occur in the former type. In relative terms, on the other hand, they are less common than omitted pronouns in declarative than in interrogative sentences, except in BrE where overt pronouns are the most frequent option in both sentence types.

Table 19. Distribution of omitted and overt pronouns per level of ‘Sentence type’

Sentence type	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Declarative	46.89% (354)	53.11% (401)	47.37% (315)	52.63% (350)	58.18% (530)	41.82% (381)
Interrogative	18.87% (10)	81.13% (43)	25.64% (10)	74.36% (29)	28.57% (10)	71.43% (25)

**Figure 20. Proportion of omitted and overt pronouns per level of ‘Sentence type’**

Focusing now on those constraints that are grounded in cognitive or processing motivations, all the hypotheses postulated in Section 5.2.3.2 are given preliminary support by the analysis. Starting with priming effects, Table 20 and Figure 21 show the distribution of

omitted and overt pronouns in the three conditions distinguished. When the previous coreferential subject in the preceding five clauses is an omitted pronoun, omission, with a relative frequency of 78.26% in BrE, 84.62% in IndE, and 90.24% in SgE, is more common than the overt expression of the subject pronoun. On the other hand, when the previous coreferential subject is realized by an overt pronoun, a noun phrase, or is not located in the preceding five clauses, overt pronouns are favoured. Omission, however, cannot be said to be completely inhibited when it is not primed, since we find a considerable number of instances in both absolute and relative terms, particularly in SgE, even when the previous coreferential subject is an overt pronoun or a noun phrase.

Table 20. Distribution of omitted and overt pronouns per level of ‘Priming’

Priming	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
NP/ No priming	37.17% (126)	62.83% (213)	41.56% (128)	58.44% (180)	53.33% (200)	46.67% (175)
Omitted pronoun	78.26% (36)	21.74% (10)	84.62% (33)	15.38% (6)	90.24% (74)	9.76% (8)
Overt pronoun	47.75% (202)	52.25% (221)	45.94% (164)	54.06% (193)	54.40% (266)	45.60% (223)

5. Complexity variance in English: Pronoun omission and language contact

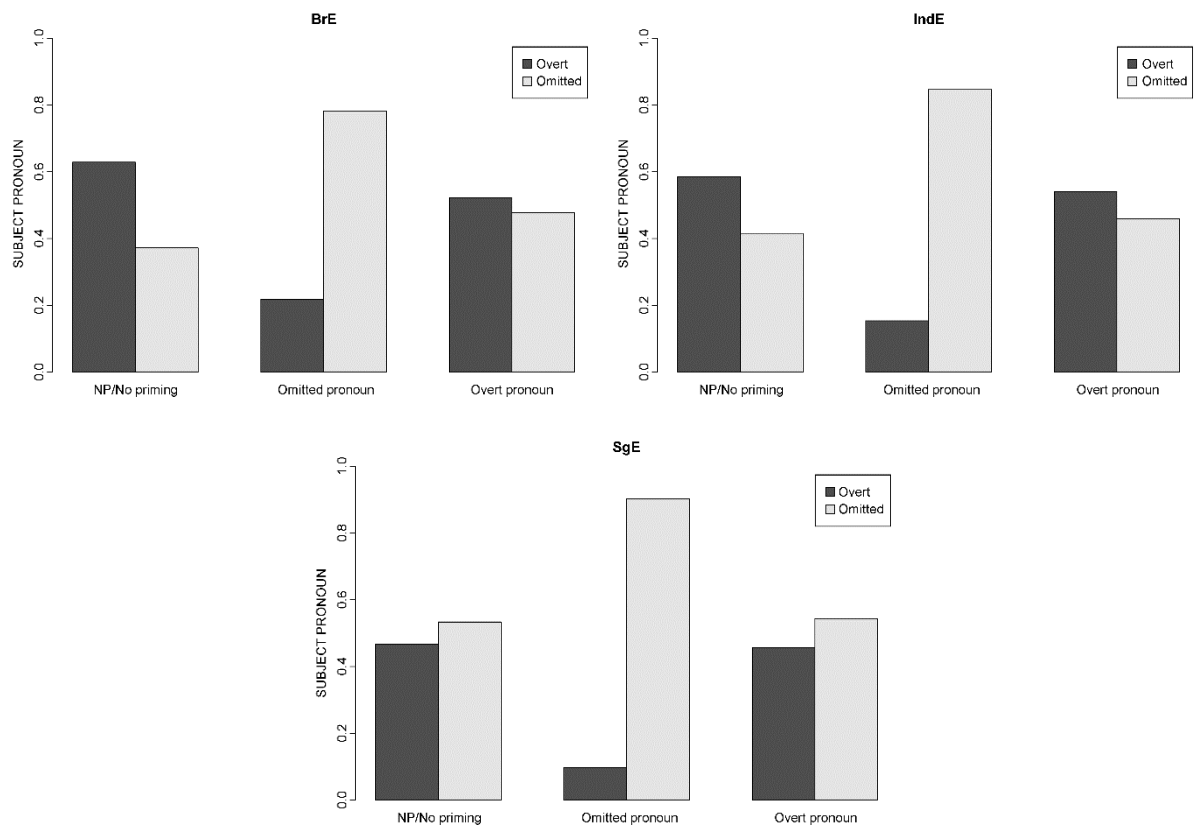
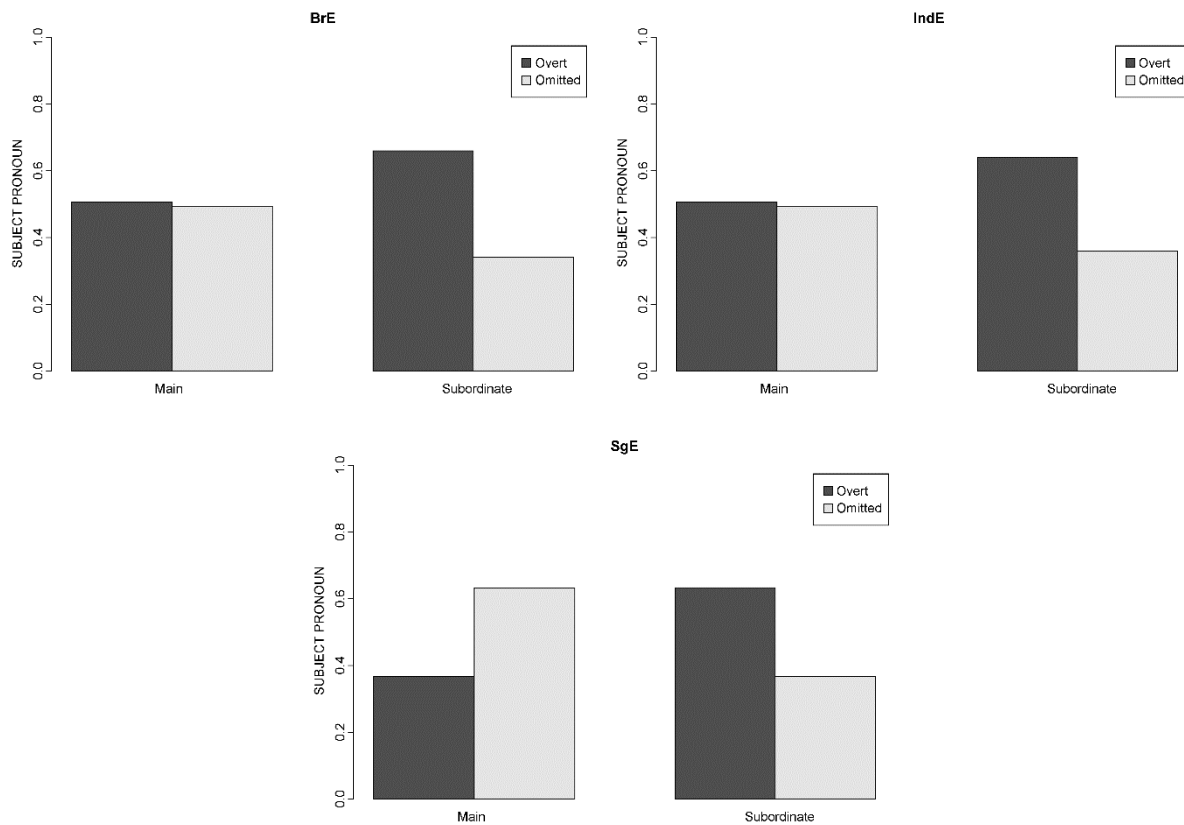


Figure 21. Proportion of omitted and overt pronouns per level of ‘Priming’

Continuing with the frequency of omitted and overt pronouns in main versus subordinate clauses, shown in Table 21 and Figure 22, omission is preferred in the former rather than the latter clause type. Even though some variation exists, speakers choose an overt pronoun when this is the subject of a subordinate clause in 65.94% of the cases in BrE, 64.12% in IndE, and 63.30% in SgE since, as discussed in Section 5.2.3.2, omitted pronouns may increase the effort needed to process this type of clauses. In main clauses, on the other hand, we find a more balanced distribution in BrE and IndE, with omitted pronouns occurring with a relative frequency of 49.40% and 45.44%, respectively. In SgE, however, omitted pronouns are clearly the preferred option in main clauses (63.19%).

Table 21. Distribution of omitted and overt pronouns per level of ‘Clause type’

Clause type	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Main	49.40% (286)	50.60% (293)	49.44% (264)	50.56% (270)	63.19% (460)	36.81% (268)
Subordinate	34.06% (78)	65.94% (151)	35.88% (61)	64.12% (109)	36.70% (80)	63.30% (138)

**Figure 22. Proportion of omitted and overt pronouns per level of ‘Clause type’**

As regards the frequency of omission of different pronominal forms, we encounter the distribution displayed in Table 22 and Figure 23. In line with Claes (2017), pronouns referring to the speaker or the hearer (or writer and reader in written language) are omitted less

frequently (38.57% in BrE, 36.01% in IndE, and 54.72% in SgE) than 3rd person pronouns (55.74% in BrE, 58.49% in IndE, and 60.88% in SgE).

Table 22. Distribution of omitted and overt pronouns per level of ‘Pronoun’

Pronoun	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Speaker/ hearer	38.57% (194)	61.43% (309)	36.01% (139)	63.99% (247)	54.72% (319)	45.28% (264)
Other	55.74% (170)	44.26% (135)	58.49% (186)	41.50% (132)	60.88% (221)	39.12% (142)

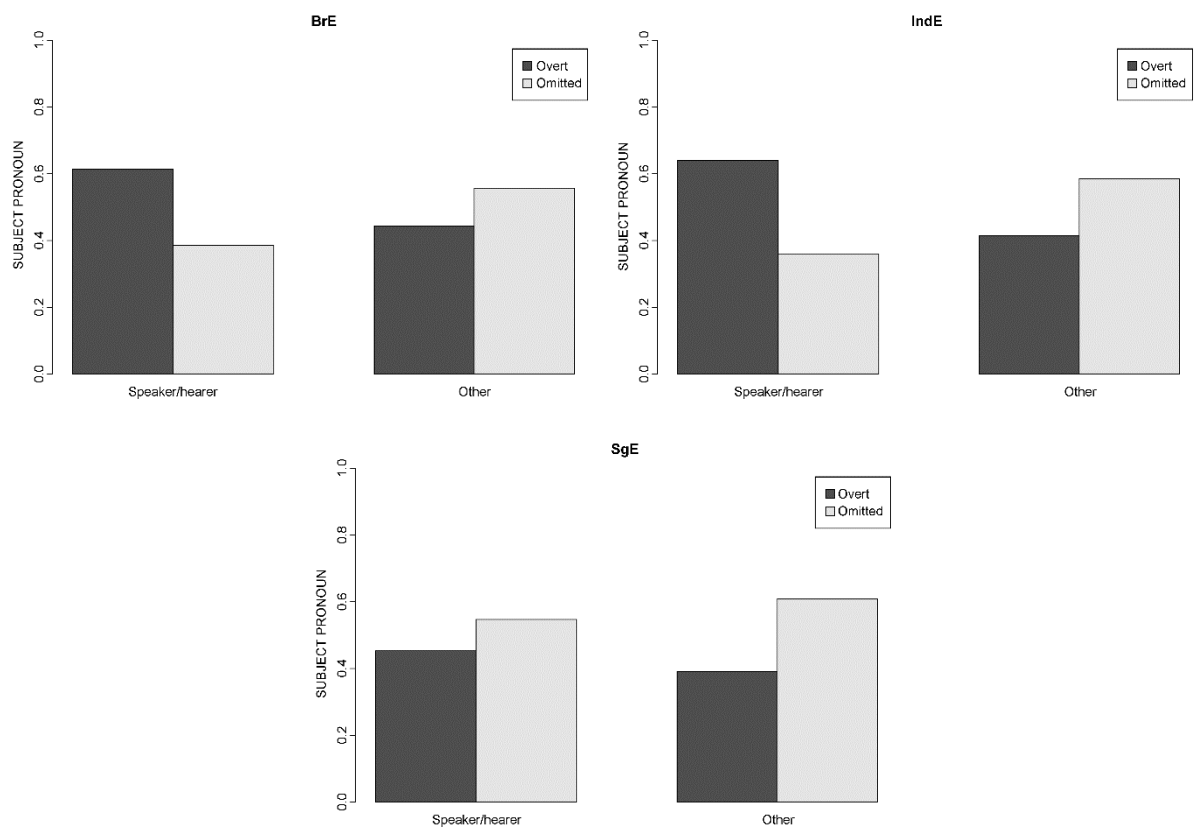


Figure 23. Proportion of omitted and overt pronouns per level of ‘Pronoun’

With respect to the numerical variables ‘Co-occurrence frequency’ and ‘Collocational strength,’ omitted and overt pronouns differ as to their median values. Overt pronouns and verbs exhibit a median co-occurrence frequency of 109.03 in BrE, 119.15 in IndE, and 137.73 in SgE, that is, pronoun-verb combinations in which the pronoun is overtly expressed co-occur a median of 137 times (per million words) in the Singaporean component of *GloWbE*. The median values observed in the case of omitted pronouns, on the other hand, are much lower than the ones of overt pronouns.

Similarly, as shown in Table 24 and Figure 25, pronouns which exhibit a stronger association with specific verb forms are usually overtly expressed, while omitted pronouns tend to occur in pronoun-verb combinations with a lower collocational strength, although the differences between omitted and overt pronouns in this respect are rather small.

Table 23. Median ‘Co-occurrence frequency’ of omitted and overt pronouns

BrE		IndE		SgE	
Omitted	Overt	Omitted	Overt	Omitted	Overt
58.71	109.03	35.56	119.15	51.49	137.73

5. Complexity variance in English: Pronoun omission and language contact

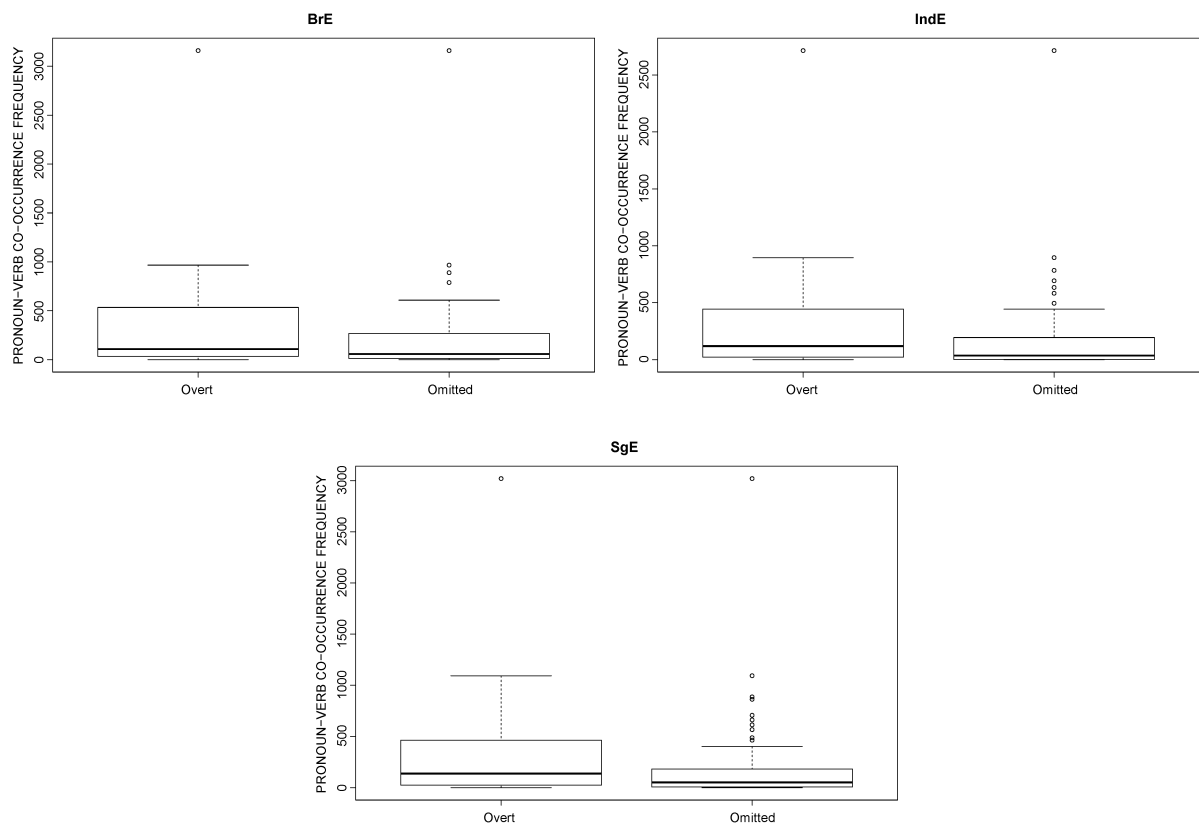


Figure 24. Median ‘Co-occurrence frequency’ of omitted and overt pronouns

Table 24. Median ‘Collocational strength’ of omitted and overt pronouns

BrE		IndE		SgE	
Omitted	Overt	Omitted	Overt	Omitted	Overt
1.29	1.45	1.32	1.40	1.31	1.42

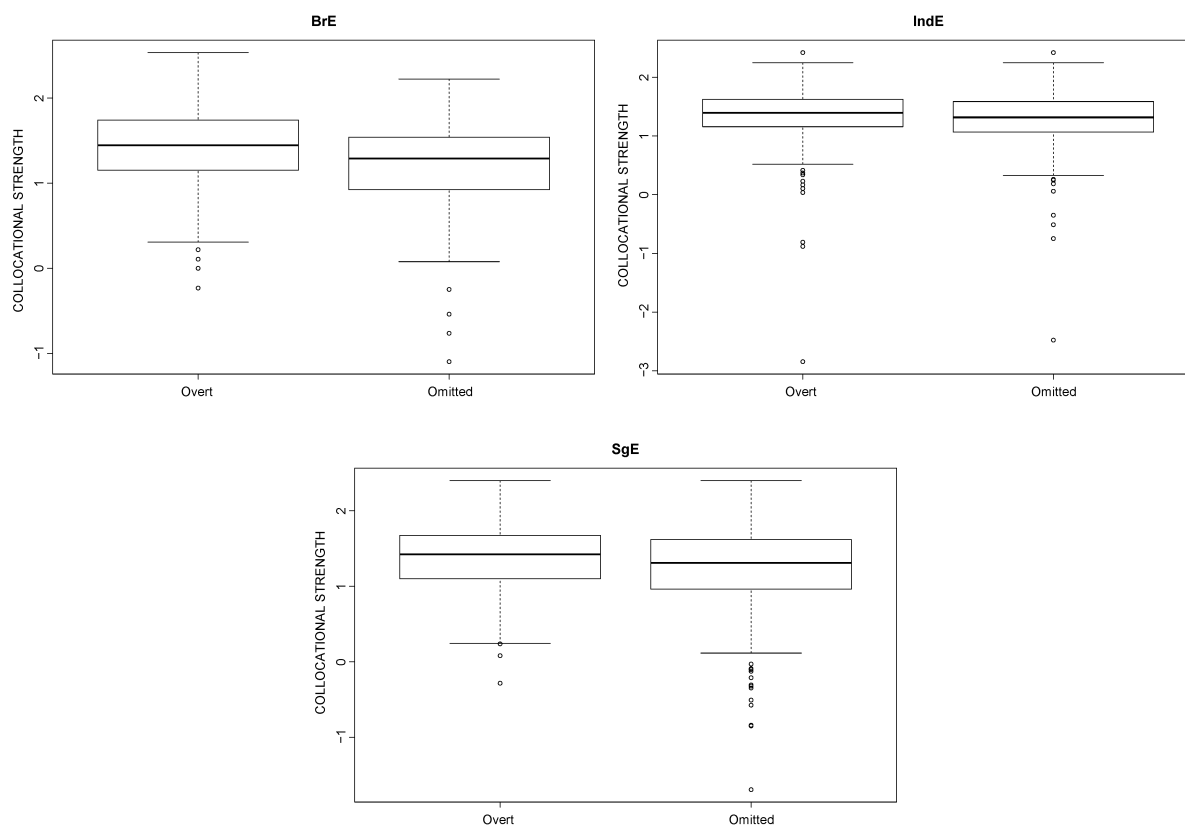
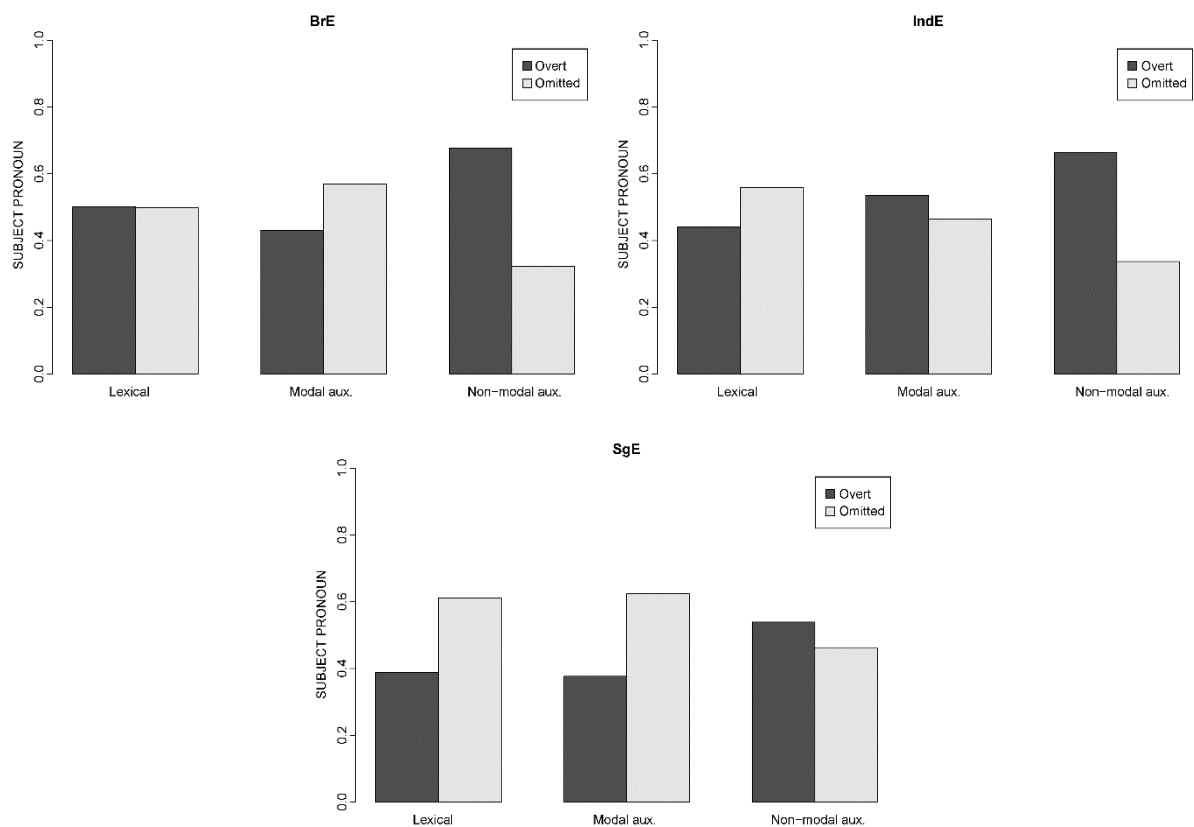


Figure 25. Median ‘Collocational strength’ of omitted and overt pronouns

Finally, frequency effects are also evident in the distribution of omitted and overt pronouns with the three different verb classes distinguished in the present study. Table 25 and Figure 26 display the absolute and relative frequencies of omitted and overt pronoun with lexical, modal auxiliary, and non-modal auxiliary verbs. As can be observed, omission is less frequent when the pronoun is followed by a non-modal auxiliary (32.29% in BrE, 33.73% in IndE, and 46.10% in SgE). Omitted pronouns, then, seem to be inhibited when they are followed by non-modal auxiliary verbs, possibly because non-modal auxiliaries and personal pronouns co-occur very frequently. Lexical verbs, on the other hand, favour omission, particularly in the high-contact varieties (49.86% in BrE, 55.91% in IndE, and 61.12% in SgE), and modal verbs increase the frequency of omitted pronouns in BrE and SgE.

Table 25. Distribution of omitted and overt pronouns per level of ‘Verb class’

Verb class	BrE		IndE		SgE	
	Omitted	Overt	Omitted	Overt	Omitted	Overt
Lexical	49.86% (176)	50.14% (177)	55.91% (175)	44.09% (138)	61.12% (305)	38.88% (194)
Modal aux.	56.87% (95)	43.11% (72)	46.48% (66)	53.52% (76)	62.36% (111)	37.64% (67)
Non-modal aux.	32.29% (93)	67.70% (195)	33.73% (84)	66.27% (165)	46.10% (124)	53.90% (145)

**Figure 26. Proportion of omitted and overt pronouns per level of ‘Verb class’**

5.3.2 Discussion

The reader may recall that the findings presented in this and the subsequent sections of this chapter are based on all the instances of omitted pronouns found in the subset of texts selected from *ICE-GB*, *ICE-IND*, and *ICE-SIN* and in an equal number of overt pronouns randomly sampled from these texts. Therefore, the percentages of omission discussed here are clearly inflated in each and every one of the contexts investigated: for instance, according to Schröter and Kortmann (2016), under no condition does the frequency of subject omission in SgE exceed 16%, and this is only in existential clauses (in other environments it is lower than 8%). This is an issue that must be borne in mind when interpreting the results. However, the focus here is not on the overall frequency of omission in BrE, IndE, and SgE but on how the different constraints influence the choice between omitted and overt pronouns, for which purpose the present data are indeed valid.

Out the three varieties examined, the only one in which omitted pronouns occur more frequently than overt pronouns is SgE, while in IndE and BrE, which behave similarly in this respect, they are less common than their overt counterparts. Therefore, structural simplification, which increases as a function of the frequency of economical structures (i.e., omitted pronouns) in a variety, is only evident in SgE. In Section 5.4.1 it will be demonstrated that there are also important, though subtler, differences between BrE and IndE in this respect, but so far it appears that these two varieties pattern closely as regards the overall proportion of omitted pronouns. This finding in turn poses a problem for the substrate influence hypothesis on the occurrence of pronoun omission features in varieties of English: the substrate languages of both IndE and SgE, as mentioned in Section 5.2.1 above, favour omission, which, other things being equal, should result in both varieties exhibiting differences with BrE. This is clearly not the case in IndE.

With regard to the other language-external constraint analysed, namely text type, a division in terms of medium of production is found, with omitted pronouns being preferred in written rather than spoken language. In addition, no clear informal versus formal split emerges from the data. These findings may at first seem surprising, as omission is characterized in reference grammars as a feature typical of casual style and/or spoken language, except when it occurs in coordinate contexts. However, omitted pronouns are the variant that provides the least information, since there is no explicit marker in the clause guiding the addressee to the correct referent. As a consequence, the process of identifying the antecedent of the omitted subject pronoun requires the addressee to have access to contextual information. In written language, the temporal constraints on spoken communication can be ignored, because the reader has all the time she or he needs to access the previous discourse and does not have the same pressure of keeping information in short-term memory (Maas 2009, 166). Therefore, the lack of an explicit marker in structures with omitted subjects may be easier to process for the addressee in written language, thus accounting for the high frequency of omitted pronouns in this type of texts.

Two contradictory hypotheses were postulated in previous research as regards the effect of subject verb agreement on the choice between omitted and overt pronouns. On the one hand, according to Cole (2009, 2010), agreement functions as an aid in order to identify the correct antecedent of an omitted pronoun, and thus the presence of agreement morphology in a particular structure should favour omission. On the other hand, Sato and Kim (2012) argued, on the basis of data from CollSgE, that omission is blocked in the presence of explicit agreement morphology. In the face of the distribution found here, none of the hypotheses is adequate to describe the present data: omission is neither favoured nor blocked in the presence of agreement. The most logical conclusion is that agreement in the varieties of

English examined here does not play a role on the choice between omitted and overt pronouns, a conclusion that will be supported by the results provided in Section 5.4.1 below. Therefore, referent accessibility emerges as the only true licenser of pronoun omission in BrE, IndE, and SgE, since deleted pronouns occur almost exclusively when they refer to highly accessible antecedents. Similarly, overt pronouns, as markers of high accessibility, are also very frequently coindexed with highly accessible referents, which support the hypothesis that accessibility does not really discriminate between omitted and overt pronouns. Its role seems to be limited to that of a necessary though not sufficient condition for the occurrence of pronoun omission. Almost no instances of deletion are coindexed with antecedents with low accessibility, and only a few cases occur with referents with intermediate accessibility. The vast majority of omitted pronouns refer back to highly accessible antecedents, which suggests that there is not a substantial increase in hidden complexity as a result of pronoun omission.

Furthermore, and as expected, omitted pronouns occur more frequently in those contexts identified in reference grammars as the canonical environments for subject omission in English, that is, in coordinate clauses when the omitted subject is located in the second or subsequent conjuncts and is coreferential with the subject of the first conjunct, in initial clause position, and in declarative sentences. However, despite their importance as explanatory factors of subject omission, many instances of omitted pronouns occur outside these contexts, which means that there is still a substantial amount of variance that is not accounted for by them.

Finally, cognitive- and processing-driven effects also emerge from the analysis. Omission is clearly more frequent when it is favoured by independently motivated cognitive or processing factors: (i) in declarative sentences, (ii) when it is primed, (iii) when it occurs in main clauses, (iv) when the pronominal form does not refer to the speaker or hearer (or writer

and reader) involved in the communicative exchange, (v) when it co-occurs with lexical and modal verbs, and (vi) when the co-occurrence frequency and collocational strength of the pronoun and the verb is relatively low. These findings provide support to the hypothesis that, when speakers omit pronouns, they do this in an efficient manner, that is, when there is not an increase in the processing load required by the addressee to successfully decode the message. This in turn means that omission indeed generally leads to structural simplification, as the resulting structures are easier for speakers to produce, by virtue of being more economical, without being harder for hearers to comprehend.

The results presented in this section are suggestive in the sense that they appear to uncover many interesting generalisations about the distribution of omitted and overt pronouns in the present data. However, they do not tell us anything about potential significant differences between the varieties with regard to effect of the other language-internal and language-external constraints. To this purpose, a multivariate analysis of the data must be carried out, which is the focus of the next section.

5.4 MULTIVARIATE ANALYSIS OF THE DATA

This section deals with the multivariate analysis of the data, that is, it presents the results of, first, the binary mixed-effects logistic regression, which are interpreted in terms of structural complexity (cf. Section 5.4.1), and, second, the per variety random forests, which are used here to measure the system complexity of the three varieties (cf. Section 5.4.2).

5.4.1 Structural complexity

5.4.1.1 Results

The final binary mixed-effects logistic regression model arrived at includes eight significant predictors and predictor interactions in the fixed effects structure, and two significant random effects. The results of the model were further validated by means of the bootstrapping method described in Section 5.2.4.1. All of the predictors turned out to be statistically significant after this validation procedure was applied, so we can be certain that the model does not include any spurious effects. The fixed significant predictors are ‘Accessibility,’ ‘Priming,’ ‘Clause type,’ ‘Pronoun,’ ‘Verb class,’ and the interactions of ‘Variety’ with ‘Text type,’ ‘Coordination,’ and ‘Clause position,’ while the two significant random predictors include ‘Speaker’ and ‘Verb form.’ Significant effects were not found for ‘Agreement,’ ‘Sentence type,’ ‘Co-occurrence frequency,’ ‘Collocational strength,’ and ‘Verb lemma.’

Table 26 reports on the goodness-of-fit statistics of the model, that is, how well the model discriminates between the two levels of the dependent variable (omitted versus overt pronouns) and how well the predictions of the model fit the observed data. These are all indications of the quality of the model achieved here.

Table 26. Model summary of the binary mixed-effects logistic regression

<i>Model summary</i>	
C-index of concordance	0.960
Correct predictions	89.91%
Pseudo- R^2 (marginal)	0.704
Pseudo- R^2 (conditional)	0.797

The second row of Table 26 contains the value of the C-index of concordance, which is a measure of how well the model discriminates between the two alternatives. This index ranges between 0.5 and 1. Values over 0.8 indicate that the model has a reasonably strong predictive

capacity and values over 0.9 reflect a model with an outstanding discriminatory power (Levshina 2015, 259). With a *C*-index of concordance value of 0.960, this is in fact the case in the present model.

The next row shows the percentage of correct predictions of the model. In Section 5.2.4.1, it was mentioned that logistic regression modelling estimates the probability of each of the levels of a dependent variable (here omitted versus overt pronouns) on the basis of a series of independent or predictor variables. That is, for each observation in the dataset the model arrives at a prediction based on the information provided by the independent variables for which the observation is annotated. These predictions are sometimes wrong: for instance, a model might predict on the basis of the predictors that a specific instance should be realized as an omitted pronoun when the observed value is in fact the opposite one, that is, an overt pronoun. Therefore, we can calculate how many times the predictions of the model fit the observed data as a measure of its predictive capacity. In this case, with a 89.91% predictive accuracy, we can be sure that we have achieved a strong model in which only 10.09% of the observations in the dataset are not correctly classified. This percentage of correct predictions is significantly better than the baseline accuracy of 50% ($p < 0.001$): since the present dataset contains an equal number of instances of omitted and overt pronouns, if one randomly guessed which level of the dependent variable occurred in each observation, one would be right approximately 50% of the times. With an accuracy of almost 90%, the model achieved here clearly surpasses this baseline.

Finally, the last two rows of Table 26 specify the marginal and conditional pseudo- R^2 values of the model, which were computed with the `r.squaredGLMM()` function in the MuMIn package (Barton 2018). Pseudo- R^2 is an additional index of predictive power and is usually equated to the proportion of variance accounted for by the model (although in logistic

regression models this is not so evident; cf. Levshina 2015, 259). Two pseudo- R^2 indexes are reported on in Table 26: the first one, marginal pseudo- R^2 , provides an indication of the predictive power of the model considering only the fixed effects, while the second one, conditional pseudo- R^2 , takes into account both fixed and random effects. What these two values reflect is that most variance explained by the model (0.704) is actually accounted for by the fixed predictors. However, by including random as well as fixed effects, the predictive power of the model increases by almost 0.1, which means that a model without random predictors would have achieved significantly worse results.

The results of the final binary mixed-effects logistic regression model are displayed in Table 27. This table is divided in two separate sections. The first one deals with the fixed effects of the model: it shows the direction and size of the effects of the fixed predictors and predictor interactions on the probability of referential subject pronoun omission, and whether these effects are statistically significant or not. The second section, on the other hand, illustrates the effects of the significant random variables. Besides being included in Table 27, the fixed and random effects will also be presented in graphical form so as to achieve a more visual representation of the model (cf. Figures 27-35).

The fixed effects section of Table 27 contains five different columns. The first column indicates the coefficients of the model. The first coefficient, called the intercept, represents the probability of omission with all the predictor variables at their reference levels. The reference levels of the predictors in the model summarized in Table 27 are BrE ('Variety'), spoken informal ('Text type'), high accessibility ('Accessibility'), no coordination ('Coordination'), non-initial position ('Clause position'), overt pronoun ('Priming'), main clauses ('Clause type'), speaker/hearer ('Pronoun'), and lexical verbs ('Verb class'). The rest of the coefficients reflect the difference in the likelihood of omission when the level of a

given predictor changes and the rest are kept at their reference values. For instance, the coefficient ‘Clause type = Subordinate’ shows the probability of omission in BrE, in spoken informal texts, with high referent accessibility, in non-coordinate contexts, in non-initial position, when the previous coreferential subject is an overt pronoun, with pronouns referring to speakers/hearers, with lexical verbs, and, crucially, in subordinate clauses. The last ten rows of the section on fixed effects in Table 25 display the significant interactions between predictors. Due to the complexity of interpreting the interaction coefficients in tabular form, they are discussed below on the basis of partial effect plots.

Table 27. Results of the binary mixed-effects logistic regression

<i>Fixed effects</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−4.239	0.562	−7.548	< 0.001
Variety = IndE	1.775	0.644	2.755	< 0.01
Variety = SgE	3.002	0.610	4.920	< 0.001
Text type = Spoken formal	−0.740	0.516	−1.435	0.151
Text type = Written informal	0.211	0.419	0.503	0.615
Text type = Written formal	0.041	0.661	0.062	0.951
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.742	0.483	11.889	< 0.001
Clause position = Initial	3.727	0.487	7.654	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.368	< 0.001
Priming = NP/No priming	0.218	0.165	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001

Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = IndE : Text type = Sp. for.	0.848	0.671	1.264	0.206
Var. = SgE : Text type = Sp. for.	0.106	0.665	0.159	0.873
Var. = IndE : Text type = Wr. inf.	−0.845	0.628	−1.345	0.179
Var. = SgE : Text type = Wr. Inf.	0.784	0.540	1.451	0.147
Var. = IndE : Text type = Wr. for.	−0.539	0.930	−0.580	0.562
Var. = SgE : Text type = Wr. for.	−0.504	0.900	−0.560	0.576
Var. = IndE : Coord. = Coord.	−0.565	0.606	−0.932	0.351
Var. = SgE : Coord. = Coord.	−1.656	0.603	−2.747	< 0.01
Var. = IndE : Cl. pos. = Initial	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Initial	−2.205	0.534	−4.127	< 0.001
Random effects				
Predictor		Variance		
Speaker		0.568		
Verb form		0.939		

The column entitled ‘Estimate’ contains the relative increases or decreases in the probability of omission in each of the coefficients as compared to the intercept. The estimate for the intercept, on the other hand, represents the probability of omission in comparison to that of overt pronouns in all the reference values of the predictors. Estimates are shown in log odds: positive values indicate a relative increase in the likelihood of omission, negative values reflect a relative decrease, and a value of zero means that there is no change.

Finally, the last three columns provide indications of the statistical significance of the estimates. Standard errors quantify the level of (un)certainly about each estimate: the higher the value, the less sure we are about a given estimate. Z-scores, shown in the third column, are calculated by dividing each estimate by its standard error, and they are used to compute the estimates' *p*-values. These are displayed in the last column, and they reveal whether the estimates are statistically significant or not.

The second section of Table 27 deals with the random effects of the model, which provide a variance value that specifies how variable the preferences of, in this case, speakers and verb forms are. Higher variances indicate stronger preferences for one of the levels of the dependent variable on the part of the individuals included in the groups. In the present model, 'Speaker' shows a lower variance (0.568) than 'Verb form' (0.939), which means that speakers exhibit weaker individual preferences than verbs for either omitted or overt pronouns.

The presence of multicollinearity in the data, that is, whether some predictors are correlated in the sense that they tap into similar underlying effects (cf. Baayen 2008, 181; Levshina 2015, 159), was also assessed. Table 28 displays the *Variance Inflation Factors* (or VIFs) of the coefficients, which represent how strongly correlated these coefficients are. As a rule of thumb, VIFs higher than five or ten are considered to indicate the presence of multicollinearity in the data (Levshina 2015, 272). Only five coefficients exceed the threshold of five ('Variety = IndE,' 'Variety = SgE,' 'Clause position = Initial', 'Variety = IndE : Clause position = Initial,' and 'Variety = SgE : Clause position = Initial'), and none of them exhibit values higher than ten. Therefore, the model presents no serious multicollinearity problems and we can safely assume that the estimate of each coefficient is not influenced by those of the rest.

Table 28. Variance Inflation Factors

Coefficient	Variance Inflation Factor
Variety = IndE	9.672
Variety = SgE	9.934
Text type = Spoken formal	4.947
Text type = Written informal	4.078
Text type = Written formal	3.869
Accessibility = Intermediate	1.167
Accessibility = Low	1.061
Coordination = Coordination	4.073
Clause position = Initial	7.887
Priming = Omitted pronoun	1.071
Priming = NP/No priming	1.230
Clause type = Subordinate	1.105
Pronoun = Other	1.303
Verb class = Modal auxiliary	1.046
Verb class = Non-modal auxiliary	1.064
Var. = IndE : Text type = Spoken formal	3.897
Var. = SgE : Text type = Spoken formal	3.484
Var. = IndE : Text type = Written informal	2.848
Var. = SgE : Text type = Written informal	3.970
Var. = IndE : Text type = Written formal	2.646
Var. = SgE : Text type = Written formal	2.988
Var. = IndE : Coordination = Coordination	2.699
Var. = SgE : Coordination = Coordination	2.784

Var. = IndE : Clause position = Initial	7.376
Var. = SgE : Clause position = Initial	8.592

In the rest of this section, the predictions of the model are individually explained on the basis of partial effect plots (cf. Section 5.2.4.1), starting with the main effects and then continuing with the interactions between ‘Variety’ and other predictors. These plots should be interpreted as follows: horizontal axes show the levels of the independent or predictor variables, while vertical axes display the probability of referential subject pronoun omission in each of the levels. In addition, whiskers represent 95% confidence intervals, that is, the range of values which, with a 95% certainty, contains the real value of the coefficient.

Figure 27 visualizes the effect of ‘Accessibility’ on the probability of encountering an omitted pronoun.

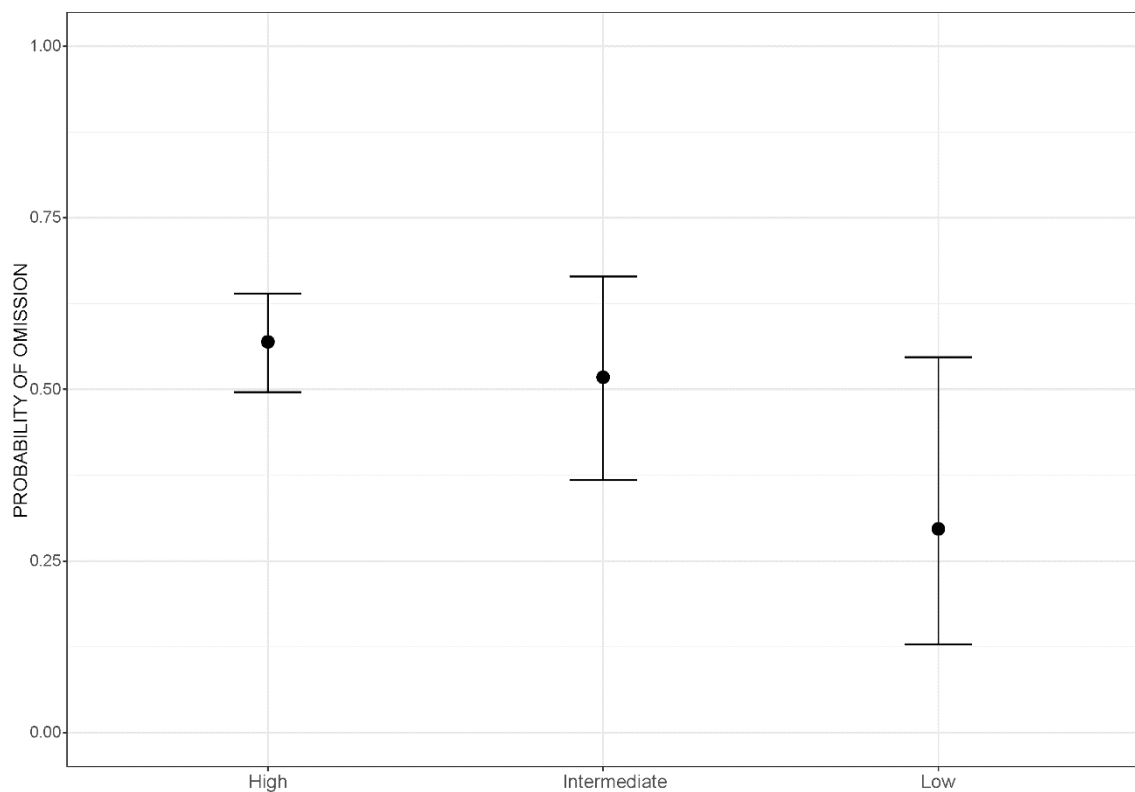


Figure 27. Probability of omission per level of ‘Accessibility’

Highly accessible antecedents favour the omission of referential subject pronouns in comparison with referents with low accessibility, and this difference is statistically significant.⁴⁹ Antecedents with intermediate accessibility occupy a middle ground between the other two levels, but they do not differ dramatically from them: neither the difference between antecedents with high and intermediate accessibility nor that between intermediate and low accessibility levels are statistically significant according to the results of the binary mixed-effects logistic regression. Despite the significant results, it seems that ‘Accessibility’ may not strongly discriminate between omitted and overt pronouns, as both referential devices are markers of high accessibility.

Priming effects are displayed in Figure 28. The preference for omitted pronouns is much stronger (and significantly so) when the previous coreferential subject in the preceding five clauses is also realized in omitted form than when this is an overt pronoun or another noun phrase. The latter two levels, on the other hand, do not differ significantly from each other.

As regards the effect of ‘Clause type,’ shown in Figure 29, it can be observed that omission is clearly favoured in main rather than subordinate clauses. This difference is statistically significant, thus agreeing with the distribution of omitted and overt pronouns found in Section 5.3.1.

⁴⁹ The significance values for most of the contrasts discussed in the present section can be found in Table 27. In other cases, however, the reference values for some predictors had to be changed and the model computed again in order to assess the significance of the relevant contrasts. These additional models can be found in Tables 45-51 in the Appendix.

5. Complexity variance in English: Pronoun omission and language contact

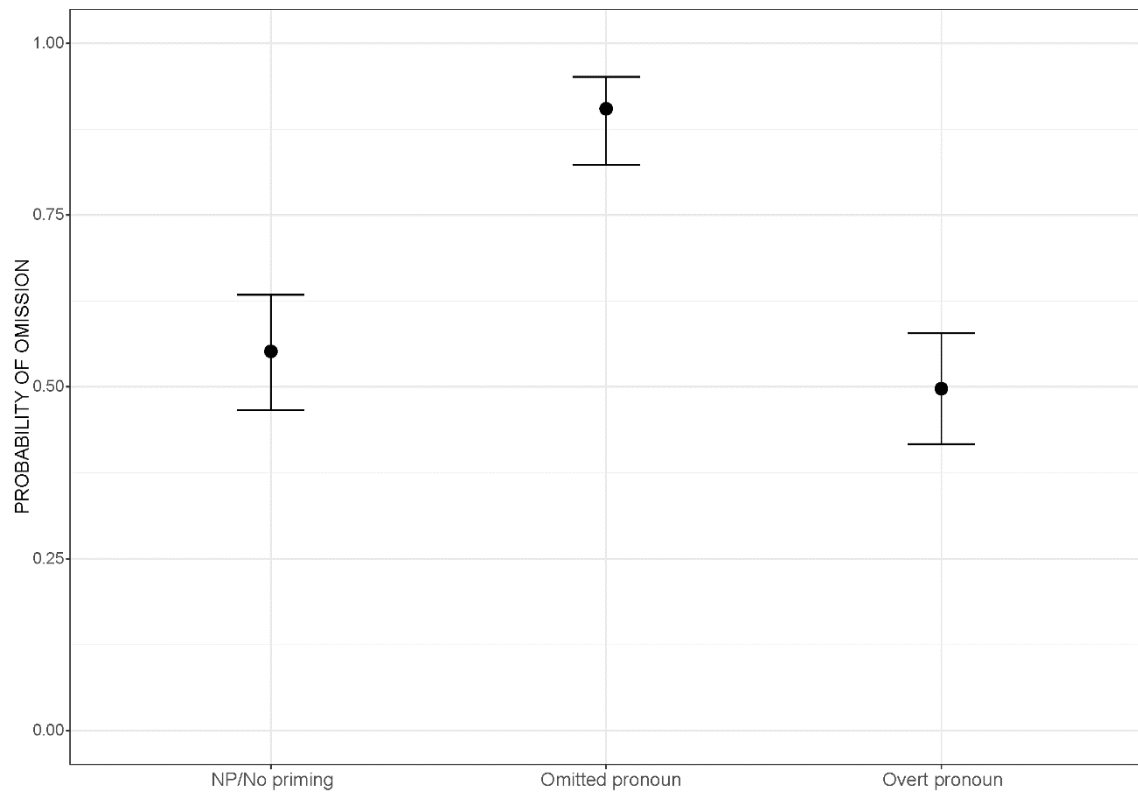


Figure 28. Probability of omission per level of 'Priming'

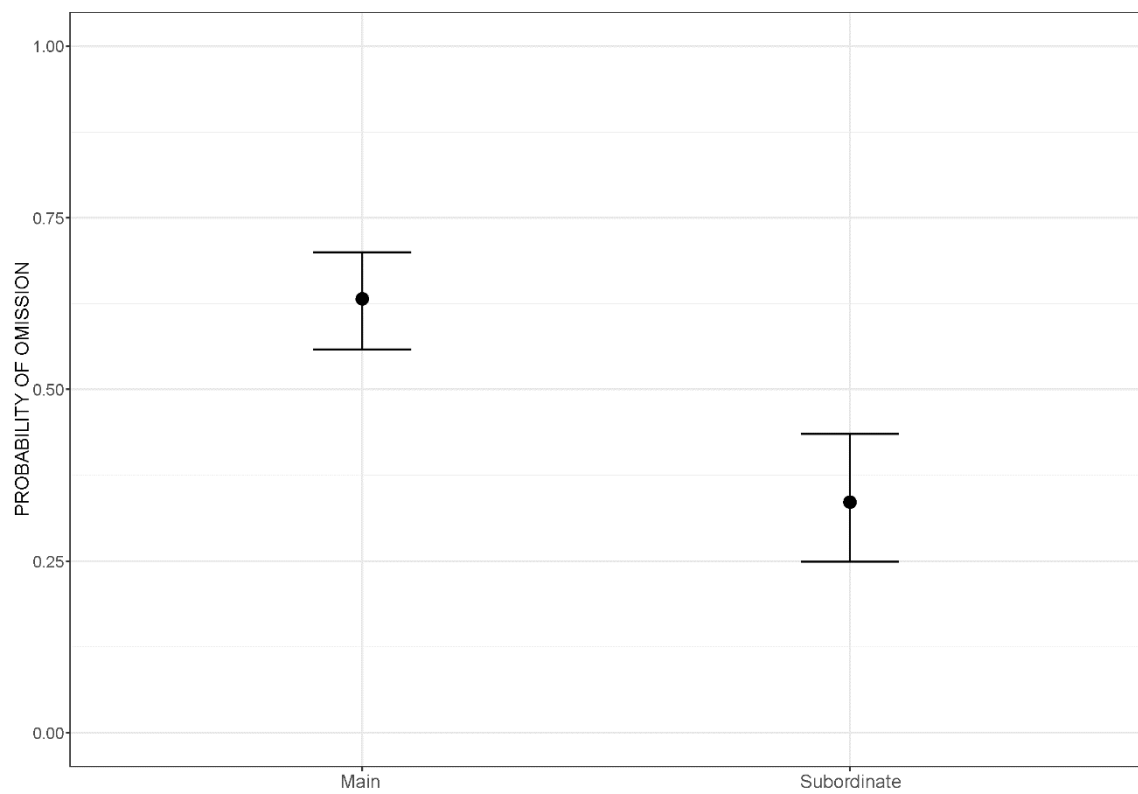


Figure 29. Probability of omission per level of 'Clause type'

With respect to the probability of omission with different pronominal forms, significant differences are found between the levels of the predictor ‘Pronoun’ (cf. Figure 30), with pronominal forms referring to speakers or hearers (or writers and readers in written texts) favour overt expression more strongly than those referring to third parties (included in the level other). The difference between the two levels, however, is not great, as was already observable in Table 22 and Figure 23 in Section 5.3.1, and we must entertain the possibility that specific pronominal forms (*I, you, she, they*, etc.) exhibit different percentages of omission. Still, the results are in line with the hypothesis postulated in Section 5.3.2, that is, omitted pronouns are found less frequently when they refer to speakers or hearers.

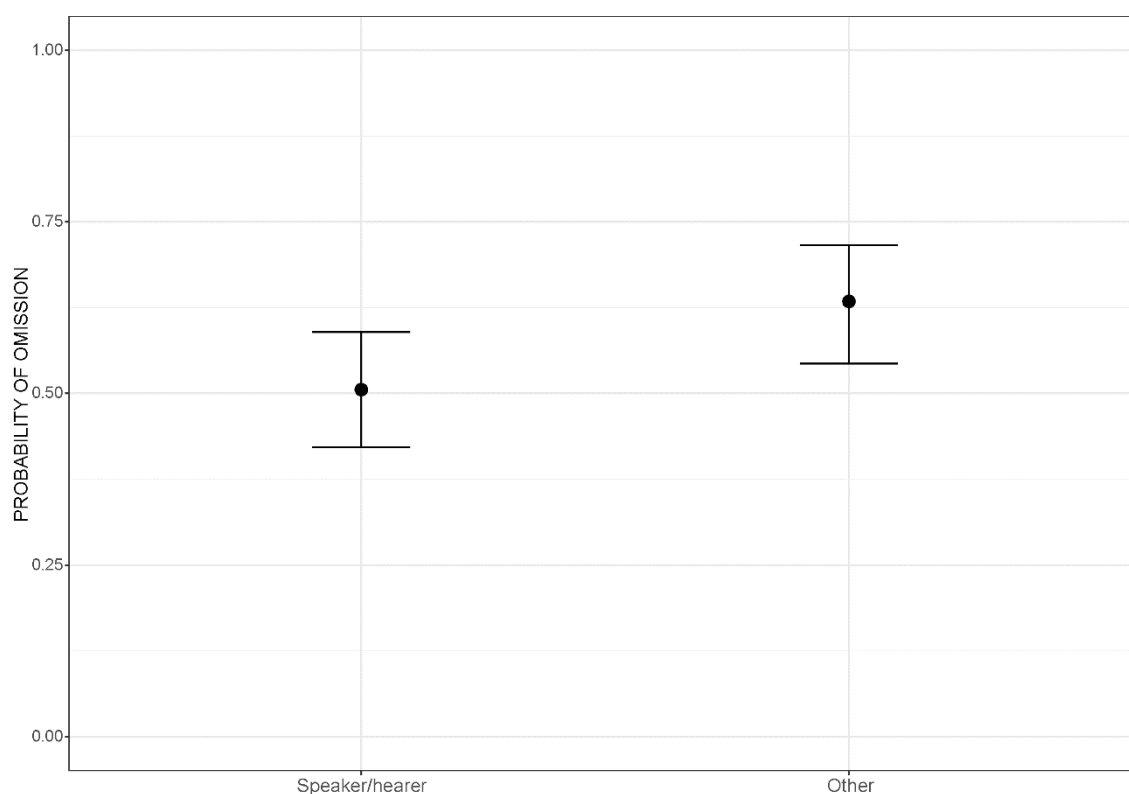


Figure 30. Probability of omission per level of ‘Pronoun’

The last main effect in the model is that of ‘Verb class,’ visualized in Figure 31. Omitted pronouns are preferred when followed by lexical and modal verbs rather than with non-modal auxiliaries; lexical verbs differ significantly from non-modal auxiliaries in this respect and modal verbs differ marginally significantly ($p = 0.072$). Despite the higher probability of omission with lexical than with modal verbs, these two verbs classes behave similarly and, in fact, no significant differences between them are found.

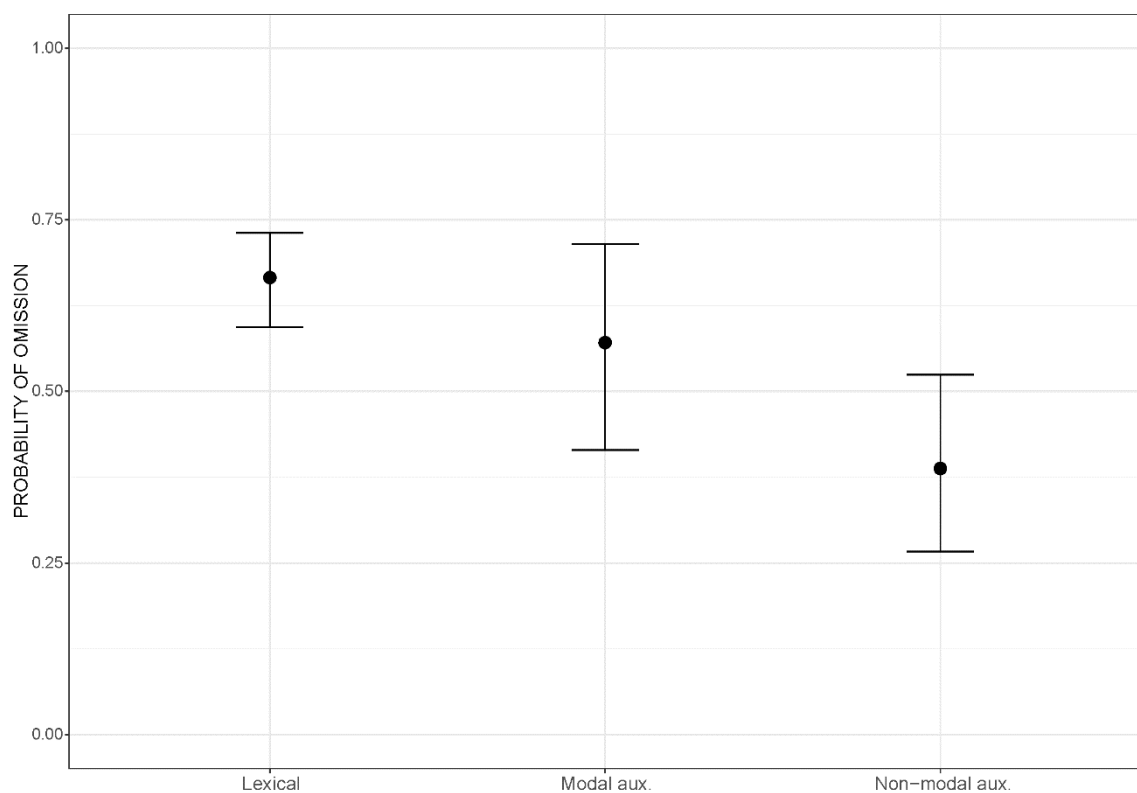


Figure 31. Probability of omission per level of ‘Verb class’

Significant interactions between ‘Variety’ and ‘Text type,’ ‘Coordination,’ and ‘Clause position,’ are found by the binary mixed-effects logistic regression, which means that the effects of these three predictors vary across BrE, IndE, and SgE. Figure 32 displays the interaction between ‘Variety’ and ‘Text type.’ In BrE, omitted pronouns are less likely in spoken formal texts (26.57%), but these only differ marginally significantly from written

informal texts ($p = 0.072$). The other three text types exhibit a probability of omission in the range of 43% to 49%. Therefore, omitted pronouns are not really favoured in BrE in any of the text types distinguished, and they are particularly dispreferred in spoken formal texts. IndE seems to present a division of labour in terms of medium of production, with spoken texts evidencing an increase in the likelihood omission (in the range of 56% to 60%) vis-à-vis written texts (41%-45%). However, no significant differences are found between the text types in this variety. Finally, in SgE, omitted pronouns are far more probable in written informal texts, with a predicted probability of 84.74%, than in the other text types, with which they differ significantly. With regard to significant differences between the three varieties, SgE outperforms both IndE and BrE in probability of omission in spoken and written informal texts, and only BrE in spoken and written formal texts. Omitted pronouns are also significantly more likely in IndE spoken informal and formal texts than in their BrE counterparts.

5. Complexity variance in English: Pronoun omission and language contact

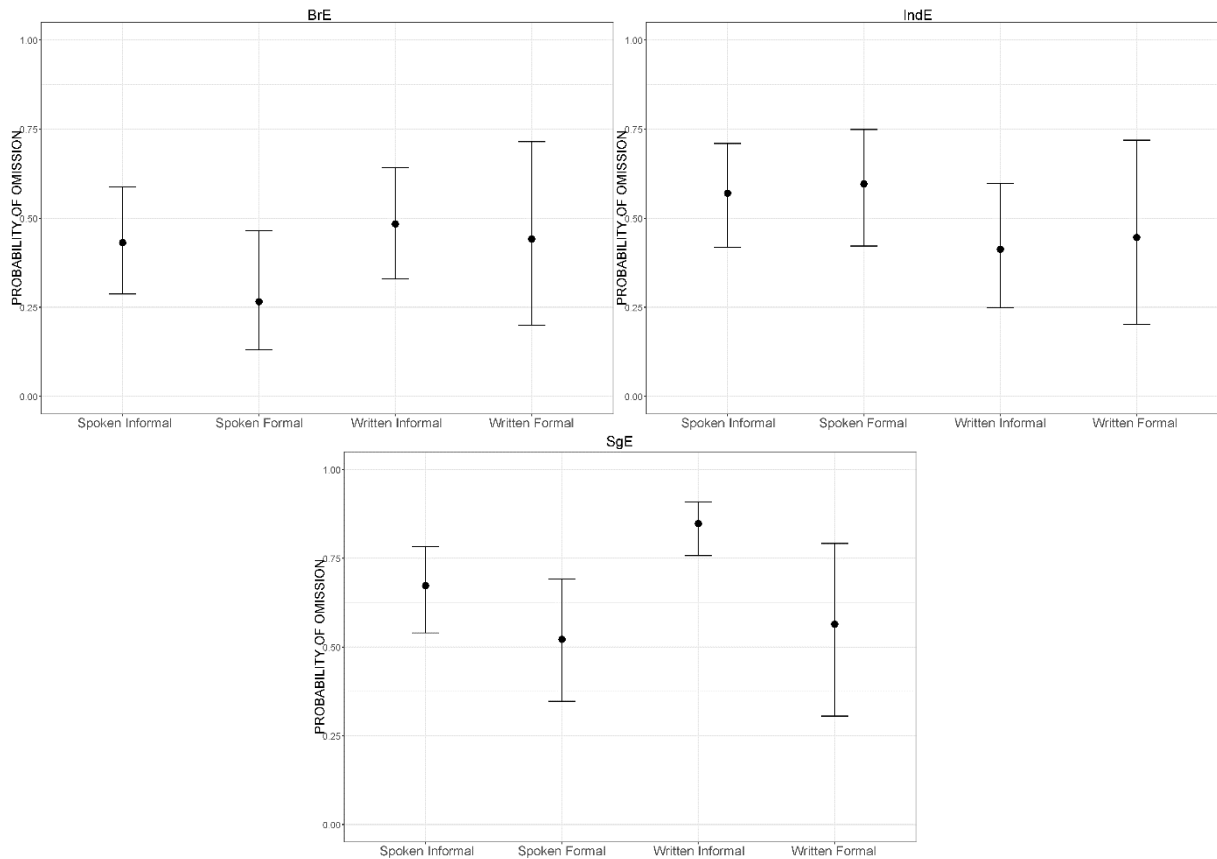


Figure 32. Probability of omission per level of ‘Text type’ in BrE, IndE, and SgE

The interaction between ‘Variety’ and ‘Coordination,’ visualized in Figure 33, reveals that omission is clearly favoured in coordinate contexts, with a predicted probability of almost 98% in the three varieties. Significant differences between the varieties, however, emerge in non-coordinate contexts. SgE exhibits a significantly higher probability of omitted pronouns in these contexts than both IndE and BrE. In addition, IndE significantly outperforms BrE in this respect, that is, it favours pronoun omission more strongly in non-coordinate contexts than BrE.

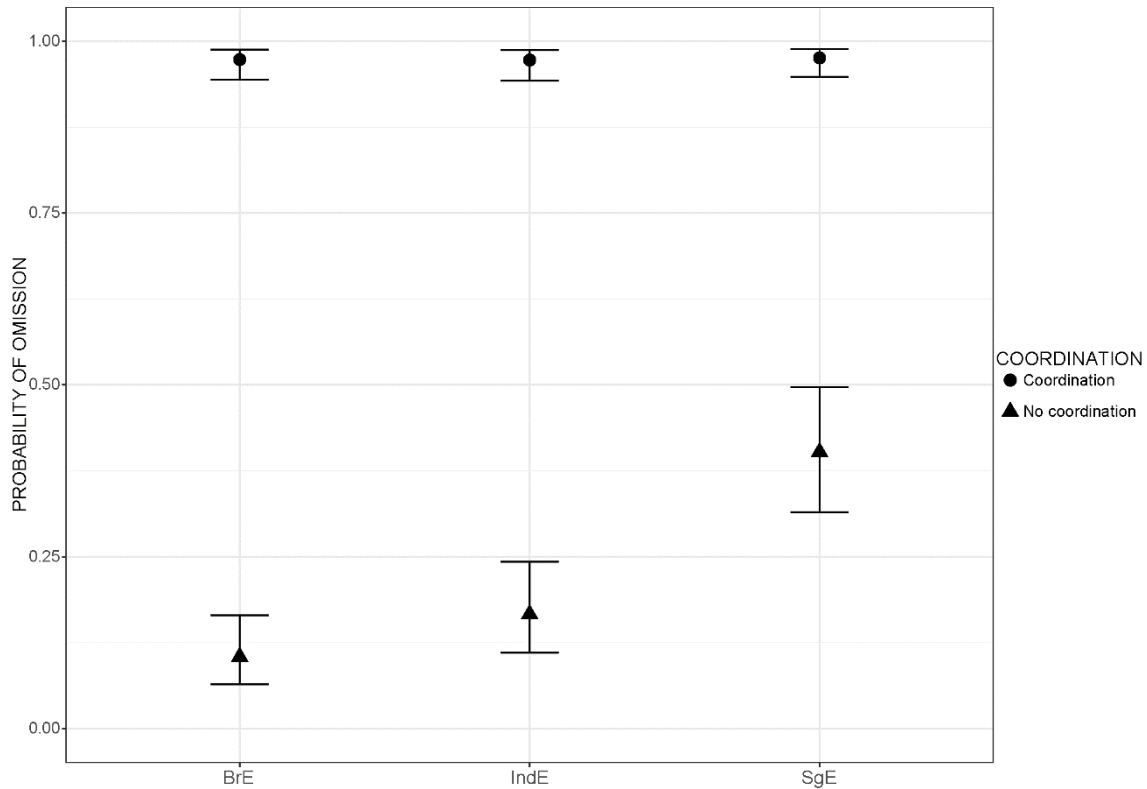


Figure 33. Probability of omission as a function of the interaction between ‘Variety’ and ‘Coordination’

A similar situation is found with respect to the interaction between ‘Variety’ and ‘Clause position,’ shown in Figure 34. Overall, omitted pronouns are preferred in initial clause position in the three varieties, with SgE displaying a significantly higher probability of omission (79.98%) than BrE (70.53%). It is in non-initial positions, however, where most differences between the three varieties emerge: omitted pronouns are significantly more likely in IndE and SgE (18.94% and 46.59% respectively) than in BrE (5.45%), and SgE also outperforms IndE in these contexts.

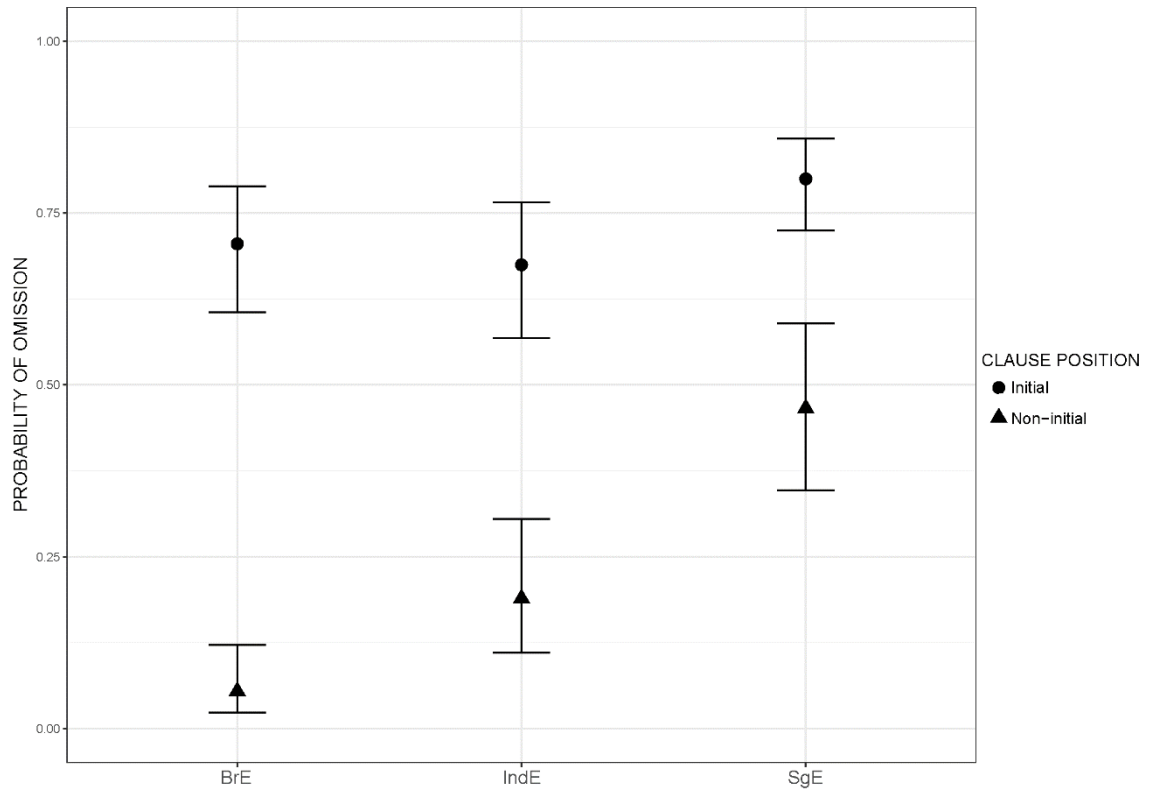


Figure 34. Probability of omission as a function of the interaction between ‘Variety’ and ‘Clause position’

As regards random predictors, Figure 35 displays the preferences for either omitted or overt pronouns of each of the verb forms in the data. Vertical axes indicate the adjustments to the intercept of each verb form: positive values reveal a preference for omitted pronouns while negative values reveal a preference for overt pronouns. Horizontal axes, on the other hand, plot the mean pronoun-verb co-occurrence frequency of each verb form, that is, the mean number of times that they co-occur with pronouns. The distributions shown in Figure 35 suggest that high-frequency pronoun-verb combinations prefer the overt expression of the pronoun. This is verified by the results of per variety Pearson product-moment correlations (cf. Levshina 2015, 116–21), which find weak but significant negative correlations between

intercept adjustments and mean co-occurrence frequencies:⁵⁰ as intercept adjustment values decrease, thus showing a stronger preference for overt pronouns, co-occurrence frequency increases, and vice versa.

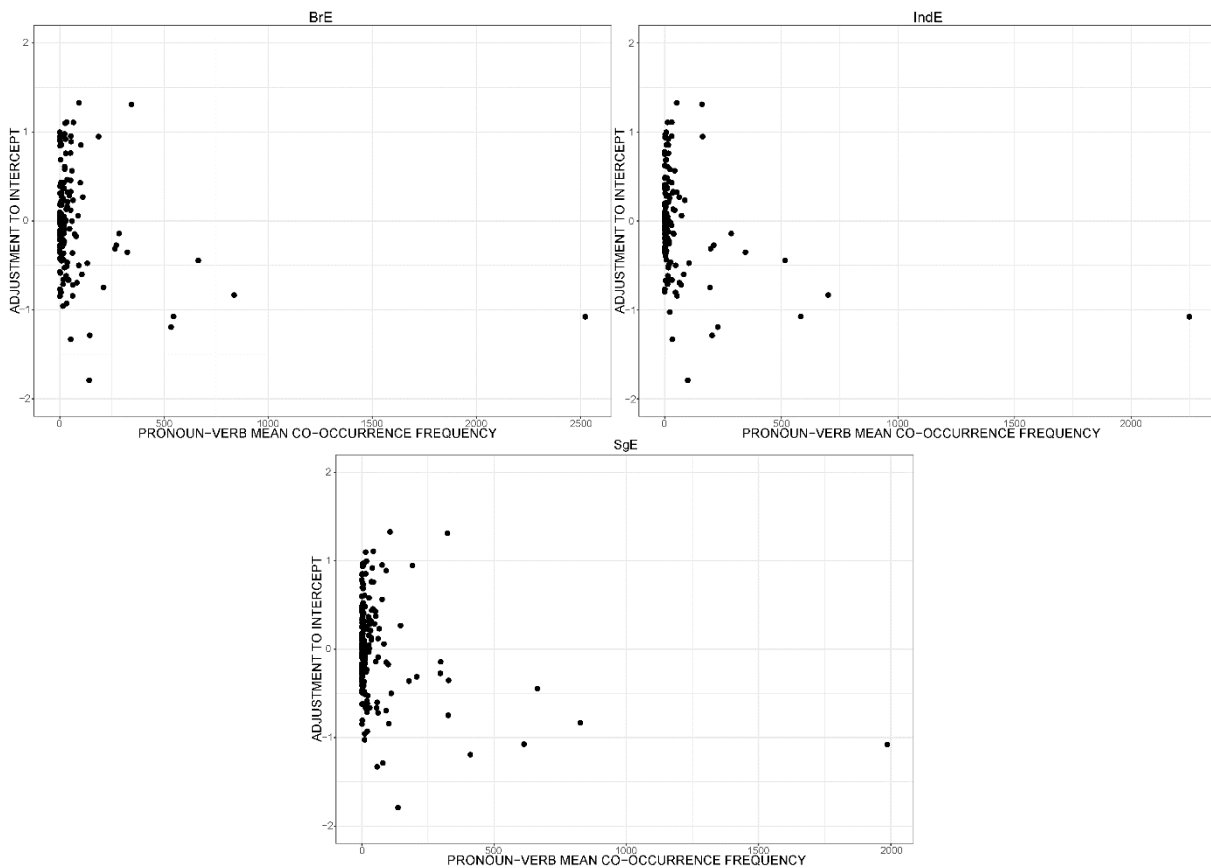


Figure 35. ‘Verb form’ effects in BrE, IndE, and SgE

Finally, the variable ‘Speaker’ was only included in the model in order to control for individual speaker preferences, as the focus of the present study is not on individual but on community-level variation patterns. Therefore, the effects of the random predictor ‘Speaker’ are not discussed here (cf. Figure 36 in the Appendix).

⁵⁰ BrE: $r = -0.244$, d.f. = 199, $p < 0.001$; IndE: $r = -0.282$, d.f. = 200, $p < 0.001$; SgE: $r = -0.242$, d.f. = 243, $p < 0.001$.

5.4.1.2 Discussion

The results of the binary mixed-effects logistic regression support many of the conclusions reached in Section 5.3. However, they also enabled us to uncover subtle differences between the varieties at hand which the previous analysis could not capture. Overall, there is an extension of pronoun omission to new contexts in SgE and IndE in comparison with BrE, even if the relative frequencies of omitted pronouns in the latter two varieties do not differ dramatically. First, referential subject pronoun omission is more probable in spoken and written informal texts in SgE than in IndE and BrE, and it is more probable in spoken informal and formal texts in IndE than in BrE. Second, SgE and IndE favour omitted pronouns in non-coordinate contexts vis-à-vis BrE, that is, outside one of the canonical contexts for omission in English, and SgE also outperforms IndE in this respect. And third, both SgE and IndE exhibit a higher probability of omission than BrE in non-initial positions, thus again in a non-canonical environment. In addition, SgE speakers favour omission more strongly than IndE speakers in non-initial clause positions. Therefore, the two high-contact varieties included in the present study display a stronger preference for omitted referential subject pronouns than BrE, the native low-contact variety, and SgE in turn displays a stronger preference than IndE. This provides further support to the claim that language contact leads to structural simplification, assuming that pronoun omission in fact results in a decrease in structural complexity (which seems to be so considering the evidence presented in previous chapters). It is suggestive that the only significant interactions found by the binary mixed-effects logistic regression are those between ‘Variety,’ on the one hand, and language-external and non-processing-driven language-internal constraints, on the other. As argued in Section 5.2.2.3, differences are expected to emerge across varieties with respect to these types of constraints: register and medium are strongly connected to social and cultural conventions, so

it is not surprising that they vary across nations, and omission in non-coordinate clauses and non-initial positions is not disfavoured by independent cognitive and processing considerations.

As regards intra-varietal differences, SgE favours omission in written informal texts. As argued in the discussion of the findings of Section 5.3.2, omission may actually be easier to process in writing than in speech, as the addressee does not have any temporal or memory limitations to identify the antecedent of the omitted pronoun. This, together with the fact that omission is characterized in reference grammars as a feature typical of casual style, except when it occurs in coordinate contexts, may explain the high probability of omitted pronouns in written informal texts in SgE. In BrE, a slight dispreference for the omitted variant was found in spoken formal texts, which may also be explained in similar terms. On the other hand, no significant differences between text types were found in IndE, which seems to point to the conclusion, also supported by the results of the per variety random forests (cf. Section 5.4.2), that ‘Text type’ does not play an important role in referential subject pronoun omission in this variety. This finding, in turn, is in line with previous characterizations of IndE as a “monostylistic variety” (Suárez-Gómez 2017, 14). On the basis of data drawn from up to ten different *ICE* components, Suárez-Gómez (2017) found that IndE exhibits a more balanced distribution of different present perfect forms across registers and modes than other varieties. The relative homogeneity of IndE, she argues, is due to the exonormative status of English in India, where speakers learn a set of prescriptive rules at school which they then put into practice in a broad range of situations.

‘Coordination’ and ‘Clause position’ have similar effects in the three varieties, notwithstanding the aforementioned differences in non-coordinate and non-initial contexts: overall, omission is more likely in coordinate clauses and initial position. The other

independent variable including a canonical context for omission in English, namely ‘Sentence type,’ was shown not to be significant by the binary mixed-effects logistic regression. This may be due to the fact that some degree of competition between omitted and overt pronouns was only observable in interrogative sentences, but this condition included a fairly low number of instances.

With respect to the main effects of the model, omitted pronouns are favoured with highly accessible antecedents; when they are primed; in main clauses; with pronouns that do not refer to speakers/hearers (or writers/readers); and with lexical and modal verbs. Consequently, pronoun deletion occurs when it is motivated by independent cognitive and processing factors and in those contexts in which it results in easier-to-process structures. These findings then provide further support to the claim made in the present book that pronoun omission leads to structural simplification, at least in relative (i.e., processing) terms. In addition, none of the processing-driven constraints interact with ‘Variety,’ which means that their effects remain stable across BrE, IndE, and SgE, possibly due to the fact that all speakers share the same cognitive architecture.

‘Agreement’ was found not to be significant by the binary mixed-effects logistic regression, which suggests that the true licenser of pronoun omission in varieties of English examined is the accessibility of the antecedent: omitted pronouns occur (almost exclusively) when they are coindexed with referents with high accessibility. The non-significance of ‘Agreement’ in the multivariate analysis is not surprising given that agreement morphology is extremely weak in English: it only unambiguously signals the person and number features of the subject in the third person present tense forms of lexical verbs, and in the forms *am* and *is* of the verb *be*.

Priming effects are observable in the regression model achieved here: omitted pronouns are more probable when the previous coreferential subject in the preceding five clauses is also realized in omitted form. When the previous coreferential subject is not an omitted pronoun, omission is not disfavoured but it is clearly less probable than when it is primed. Similarly, omission is more likely in main clauses, in accordance with Rohdenburg's (1996) cognitive complexity principle, which states that less transparent variants are dispreferred in cognitively complex environments, such as in subordinate clauses. In addition, in line with Claes (2017), deleted pronouns are chosen over overt pronouns when they refer to entities that are not the interlocutors of the conversation (or the writer and reader of a written text). The difference in the likelihood of omission between pronominal forms that refer to speakers/hearers and those that refer to other entities is not great but, overall, the hypothesis seems to hold. Finally, omitted pronouns appear to be inhibited when they are followed by non-modal auxiliaries. As shown in Section 5.2.2.2, non-modal auxiliaries and personal pronouns co-occur very frequently, meaning that certain pronoun-auxiliary combinations may be stored in memory as single units. This may in turn discourage the omission of the pronoun, as the easiest option may be to retrieve and produce these units as wholes with the verb occurring in a contracted form.

Frequency effects are only included in the model by means of the categorical variable 'Verb class.' The numerical predictors 'Co-occurrence frequency' and 'Collocational strength' turned out not to be statistically significant in the multivariate analysis of the data. This may in part be due to the fact that frequency effects are already captured, to a certain extent, by the random effect of 'Verb form': as shown in Figure 34, verb forms with a high mean pronoun-verb co-occurrence frequency tend to favour the overt expression of the pronoun.

From a structural perspective, therefore, omission results in simpler and efficient structures, since omitted pronouns occur more frequently in contexts in which they are favoured by independent cognitive and processing motivations. High-contact varieties exhibit a preference for omission vis-à-vis BrE: omitted pronouns increase in probability in certain contexts in which they are not favoured in the native variety. SgE and IndE are thus structurally simpler than BrE in this respect. However, the system complexity of the pronoun-deletion grammars of the three varieties have not so far been discussed. This is the focus of the next section.

5.4.2 System complexity

5.4.2.1 Results

Per variety random forests were computed in order to measure the varieties' system complexities. Only those predictors that turned out to be significant were included in this step of the analysis. In addition, 'Speaker' was not considered in the random forests as we are not interested at this point in the importance of individual preferences but rather in that of the more purely linguistic constraints. Table 29 contains summary statistics of the three random forests. As can be observed, the three per variety models exhibit extremely high *C*-index values and classification accuracies: the *C*-index is higher than 0.95 in the three cases, and over 86% of the predictions fit the observed data in the three varieties.⁵¹ These values indicate that the three models achieved here have a strong predictive capacity and an outstanding discriminating power.

⁵¹ In all three models, the percentage of correct predictions is a significant improvement over the baseline of 50% ($p < 0.001$ in the three cases).

Table 29. Model summaries of BrE, IndE, and SgE random forests

Model summary: BrE	
C-index of concordance	0.976
Correct predictions	90.93%
Model summary: IndE	
C-index of concordance	0.963
Correct predictions	89.54%
Model summary: SgE	
C-index of concordance	0.953
Correct predictions	86.85%

The results of the per variety random forests are displayed in Table 30 and Figure 36. In both cases the relative importance values indicate how well each of the variables considered discriminates between the omitted and overt referential subject pronoun variants. In Figure 36, the constraints are ordered as to their relative importance in each variety. Vertical dashed lines signal the absolute values of the variables with the smallest score in each variety, which, as mentioned in Section 5.2.3.2, designate the threshold of significance: only those variables that deviate from those values are considered to be relevant.

The first issue to consider is the number of variables that can be said to play some role in the choice between omitted and overt pronouns in each variety. A visual inspection of Figure 36 suggests that, in BrE, the first seven variables (‘Coordination,’ ‘Clause position,’ ‘Verb form,’ ‘Text type,’ ‘Clause type,’ ‘Priming’ and ‘Verb class’) have a significant relative importance. ‘Pronoun,’ on the other hand, plays only a extremely minor role and ‘Accessibility’ is clearly not an important constraint in BrE. In IndE, the first seven predictors

can be considered to discriminate between omitted and overt pronouns; these are ‘Coordination,’ ‘Clause position,’ ‘Verb form,’ ‘Verb class,’ ‘Pronoun,’ ‘Accessibility’ and ‘Text type,’ even though the role of this last constraint is very minor. The final two variables, that is, ‘Clause type’ and ‘Priming,’ are not really relevant in the pronoun-deletion grammar of IndE. Finally, SgE is the variety with the largest number of important constraints: up to eight variables play an important role in its pronoun-deletion grammar, and only ‘Accessibility’ can be said to be superfluous.

Table 30. Variable importance in BrE, IndE, and SgE

Predictor	BrE	IndE	SgE
Text type	0.01472	0.00747	0.03012
Accessibility	0.00299	0.00812	0.00219
Coordination	0.20794	0.20079	0.12737
Clause position	0.04829	0.04737	0.46361
Priming	0.01062	0.00438	0.01086
Clause type	0.01113	0.00455	0.02334
Pronoun	0.00484	0.01107	0.02156
Verb class	0.01017	0.01577	0.00596
Verb form	0.02452	0.02487	0.04119

With regard to the relative ranking of constraints in each variety, it is worth noting that ‘Coordination’ is ranked first in all three varieties. This is one of the canonical contexts for omission in English identified in reference grammars, and it is also the most important constraint in the three varieties examined here. ‘Clause position,’ another canonical omission environment, also ranks second in the three varieties, which means that the position of the

referential expression in the clause has a strong discriminatory power between omitted and overt pronouns. The ranking of ‘Verb form’ effects in the per variety random forests, third in all three varieties, highlights the importance of taking lexical effects into consideration. In the present study, the individual preferences of the verb forms found in the data play a much more significant role in the choice between omitted and overt pronouns than many other constraints traditionally examined in previous research on pronoun omission, such as ‘Priming’ or ‘Accessibility.’

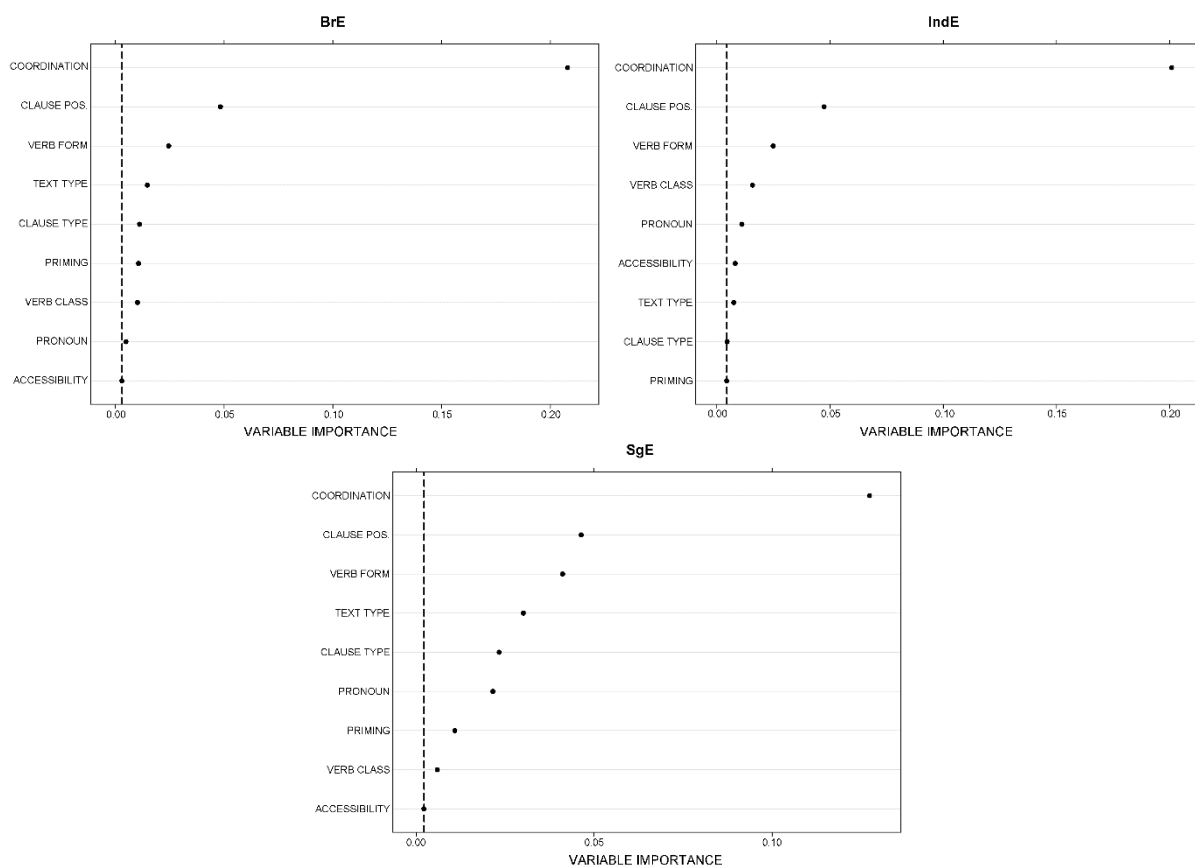


Figure 36. Variable importance in BrE, IndE, and SgE

IndE starts to diverge from BrE and SgE, which exhibit a more similar ordering of constraints, from the fourth position in the ranking downwards. ‘Text type’ is found in fourth

position in BrE and SgE but it is relegated to the seventh place in IndE. Similarly, ‘Clause type’ ranks fifth in both BrE and SgE, but only eighth in IndE. ‘Pronoun,’ ‘Priming,’ ‘Verb class,’ and ‘Accessibility’ occupy the last four positions in BrE and SgE (although not in exactly the same order), with ‘Accessibility’ being last in both varieties. In IndE, on the other hand, ‘Priming’ occupies the last position in the ranking, while ‘Verb class,’ ‘Pronoun,’ and ‘Accessibility’ are found in the fourth, fifth, and sixth positions respectively.

5.4.2.2 Discussion

The results of the per variety random forests confirm the hypothesis postulated in Section 5.1 about the system complexities of the three varieties. The number of clearly important predictors goes from eight in SgE, to seven in BrE and IndE. These values indicate the number of language-internal and -external constraints that the speakers of each variety need to take into consideration when (consciously or unconsciously) choosing between producing a referential subject pronoun in omitted or overt form. Consequently, the larger this number of constraints, the more systemically complex the pronoun-deletion grammar of the variety, and a more complex grammar is also harder to learn. IndE, being the only ‘pure’ second-language variety in the set, exhibits the simplest pronoun-deletion grammar, with only seven important constraints. SgE, which lies somewhere between BrE and IndE in terms of ‘nativehood,’ surpasses the other two varieties with respect to its system complexity, with eight constraints influencing the choice between omitted and overt pronouns. Finally, BrE occupies an intermediate position between SgE and IndE in terms of system complexity, since it exhibits seven important constraints and one predictor, namely ‘Pronoun,’ which plays a very minor but still significant role in its pronoun-deletion grammar. The reader may recall that Trudgill (2011; cf. Chapter 2, Section 2.2.3) proposed a typology of contact situations with different

effects on the complexity of languages/dialects. He argued that situations involving intensive contact over a long period of time and, crucially, extensive childhood bilingualism lead to grammatical complexification due to additive borrowing and contact-induced grammaticalization. This is the type of contact situation that best describes the current state of affairs in Singapore, and the increase in the system complexity of the pronoun-deletion grammar of SgE vis-à-vis BrE and IndE may be understood in these terms.

As regards the relative ranking of constraints, ‘Coordination,’ ‘Clause position,’ and ‘Verb form’ are the most important predictors of pronoun omission in the three varieties, which suggests a reasonably high degree of grammatical homogeneity between them. The first two are canonical contexts for omission in English, as mentioned in reference grammars, so their importance in the varieties examined is not surprising. The position of ‘Verb form’ in the ranking agrees with the findings of Torres Cacoullos and Travis (2014; cf. Chapter 3, Section 3.3.2). They argued that three partially lexically filled constructions, repeated here in (64) for convenience, account for the phenomenon of 1st person singular subject omission in spoken AmE:

- (64) a. Coordination, [*I_i* VERB *and* *Ø_i* VERB]: **I_i** went and **Ø_i** got a wet rag and wiped it off the car (Torres Cacoullos and Travis 2014, 26)
- b. IU initial position, [*Ø* VERB ...]_{IU}: **Ø** Guess they gotta make money somehow (Torres Cacoullos and Travis 2014, 23)
- c. Persistence (both within and outside coordination contexts), [*Ø_i* VERB (*and*) *Ø_i* VERB]: last Sunday, **Ø_i** got there at eight, **Ø_i** left at ten, **Ø_i** dropped this person off at home (adapted from Torres Cacoullos and Travis 2014, 29)

The findings of the present study suggest that we may benefit from extending this approach to other pronominal forms and other varieties.

‘Text type’ and ‘Clause type’ are also important constraints in BrE and SgE, and they hold the same relative positions in these varieties. In IndE, on the other hand, they are relegated to the eighth and seventh positions respectively.⁵² Similarly, the lowest positions in the ranking were occupied by the same four constraints in BrE and SgE (i.e., ‘Pronoun,’ ‘Priming,’ ‘Verb class,’ and ‘Accessibility’), while only two of these four predictors were found in comparable ranks in IndE. This seems to suggest that BrE and SgE are more homogeneous with respect to their pronoun-deletion grammars than either of them is with IndE.

Finally, the position of ‘Accessibility’ in the varieties’ pronoun-deletion grammars is noteworthy: it does not play a role at all in BrE and SgE, and only a minor one in IndE. As argued in Section 5.3.2, referent accessibility seems to be a necessary though not a sufficient condition for the occurrence of omitted pronouns, that is, they must be coindexed with highly accessible antecedents, but this is not enough. Accessibility does not really seem to discriminate between omitted and overt pronouns, as both referential expressions are commonly used to refer to highly accessible entities or events in the discourse.

5.5 COMPLEXITY AND CONTACT REVISITED

The results of the present study suggest the existence of an intricate relation between language contact and complexity, at least with respect to pronoun omission. Overall, omitted pronouns are more frequent in SgE than in IndE and BrE, which display very similar percentages of

⁵² The low importance of ‘Text type’ in IndE was already evident in the results of the binary mixed-effects logistic regression, since no significant differences between text types in this variety emerged from the analysis.

omission. Further differences between the varieties emerge with regard to the effects of other constraints, differences which can be characterized as extensions of omission to new contexts of use in the high-contact varieties. First, deleted pronouns are favoured in spoken and written informal language in SgE vis-à-vis IndE and BrE, and in informal and formal conversation in IndE as compared to BrE. Second, SgE and IndE exhibit a stronger preference for the omitted variant in non-coordinate contexts than BrE, and SgE a stronger preference than IndE. And third, both SgE and IndE speakers omit pronouns more frequently in non-initial clause position than BrE users, with SgE speakers also outperforming IndE speakers in this respect. The other language-internal constraints behave similarly in the three varieties, with pronoun deletion being favoured with highly accessible antecedents, when it is primed, in main clauses, with 3rd person pronouns, with lexical and modal verbs, and in low-frequency pronoun-verb combinations.

Omitted pronouns clearly represent a structural simplification in comparison with overt pronouns: the former referential expressions occur mostly when they result in economical structures without a concomitant loss in clarity (and, consequently, without a substantial increase in hidden complexity), as they almost exclusively refer to highly accessible antecedents. Hawkins (2014, 34; cf. Chapter 2, Section 2.3.3) argues that:⁵³

Communication is efficient when the message intended by S is delivered to H in rapid time and with the most minimal processing effort that can achieve this communicative goal.

⁵³ S and H stand for ‘Speaker’ and ‘Hearer’ respectively.

Therefore, omission results in more efficient communicative acts. The message is transmitted faster and with less processing effort in structures with deleted subjects than in those with overt subjects, as there is an extra form in the latter type of structures that must be articulated and processed. In addition, as shown by the results of the present study, speakers only delete referential subject pronouns when they consider that their antecedents are highly accessible for addressees and in those structural contexts in which omission does not lead to processing difficulties.

The varieties also differ at the systemic level. The number of important constraints is larger in SgE, with eight factors influencing the choice between omitted and overt pronouns, than in BrE and IndE, with seven clearly important constraints. Moreover, BrE exhibits an additional relatively important predictor, ‘Pronoun,’ which plays a role, even though a very minor one, in its pronoun-deletion grammar. ‘Coordination,’ ‘Clause position,’ and ‘Verb form’ emerge as the most important variables in all three varieties, thus suggesting a certain degree of grammatical homogeneity across them. However, IndE diverges dramatically from SgE and BrE, which display rather similar pronoun-deletion grammars, with respect to the importance of the rest of the constraints.

If we interpret these findings in terms of the complexity of the varieties, the following clines can be postulated, where ‘>’ means ‘more complex than’:

- Structural complexity: BrE > IndE > SgE
- System complexity: SgE > BrE > IndE

The two high-contact varieties examined here are thus structurally simpler than BrE, with SgE being simpler than IndE due to its high relative frequency of omission and its stronger

preference for omission in informal language, non-coordinate contexts and non-initial positions. It seems, then, that contact fosters structural simplification. In systemic terms, on the other hand, IndE is simpler than BrE and SgE, as more constraints are necessary to account for the variation between omitted and overt pronouns in the latter two varieties, while SgE exhibits the most complex pronoun-deletion grammar. Second-language acquisition, therefore, fosters systemic simplification, since IndE is the only pure second-language variety of the set: BrE is clearly a native variety, and SgE is in a process of shift from second to first language.

Overall, BrE is the most complex variety of the ones examined here, since it is the most structurally complex one and the second most complex in systemic terms. Contrariwise, IndE is the simplest variety: it is the simplest systemically speaking and the second variety in the structural complexity cline. The situation with respect to SgE is less straightforward, as it is the simplest variety in structural terms but the one with the most complex pronoun-deletion grammar. This may in turn point to a trade-off between structural and system complexity in this variety (and possibly also in BrE and IndE, although less clearly so). It could thus be the case that the increase in frequency of pronoun omission and its extension to new contexts may require the support of a more complex grammar, that is, a larger number of constraints establishing when to use each of the alternatives.

Finally, the results achieved in the present study suggest that substrate influence does not account for many of the differences found between BrE, IndE, and SgE as regards the deletion of referential subject pronouns. For one, IndE does not differ from BrE in the relative frequency of omitted pronouns, as both varieties exhibit highly comparable omission percentages, and this is so despite the fact that IndE has an omission-friendly substrate. Moreover, SgE, even though it displays a higher frequency of omission, does not much differ

from BrE in its pronoun-deletion grammar. In fact, the rankings of constraints as to their relative importance in these two varieties are highly correlated, as shown by the results of a Spearman's rank-order correlation (cf. Levshina 2015, 130–33).⁵⁴ If the substrate languages of the varieties played an important role in the omission of subject pronouns, we would expect to find more differences between BrE, on the one hand, and IndE and SgE, on the other. Instead, each of them resembles BrE in particular aspects that are better explained by their statuses as second- or first-language varieties.

⁵⁴ $\rho = 0.95$, d.f. = 7, $p < 0.001$.

6 CONCLUDING REMARKS AND SUGGESTIONS FOR FURTHER RESEARCH

Language complexity and communicative efficiency are two topics with a long research tradition in linguistics. Despite this, I believe that this monograph has offered new and interesting findings by applying these notions to the study of pronoun omission in different varieties of English. The overarching aim of the present work was to assess the contribution of pronoun omission to the complexity and efficiency of the grammars of several varieties of English. The conclusion was reached that it generates simpler structures, which are also more efficient provided that the antecedents of omitted pronouns are highly accessible to the addressee. That is, in terms of structural complexity, pronoun omission results in simplification. This claim was first supported on theoretical grounds in Chapters 2 and 3, by applying several complexity metrics and notions postulated in previous research, and then on empirical grounds in Chapters 4 and 5. On the other hand, from a systemic perspective, pronoun omission increases the complexity of the grammar in tectogrammatical terms, as it necessarily entails a larger set of referential expressions, and thus of structural variants. In addition, it may also result in complexification in phenogrammatical terms if more rules or constraints are needed to regulate the choice between the different alternative referential expressions.

The second goal of this book was to provide further support to the well-established claim that language contact leads to simplification. Not all kinds of contact, however, culminate in simpler grammars. As mentioned in Chapter 2, Section 2.2.3, Trudgill (2011) postulated a typology of contact situations and their effects on the complexity of languages. Three types of situations were put forward by Trudgill. Short-term widespread second-language acquisition and use by adolescents and adults foster grammatical simplification as a result of imperfect learning, among other factors. Intensive contact over a long period of time involving extensive childhood bilingualism, on the other hand, promotes complexification due to, for instance, the borrowing of grammatical distinctions or the grammaticalization of new categories in one language on the basis of another. Lastly, low-contact isolated communities preserve or even increase grammatical complexity due to the spontaneous growth of grammatical categories. Three varieties of English, namely BrE, IndE, and SgE, roughly corresponding to these three distinct types of contact situations, were selected and analysed in Chapter 5 to evaluate the usefulness of Trudgill's typology in order to account for differences in their degree of structural and system complexity. The results presented in this chapter showed that the varieties differed in terms of their complexity, at least with respect to pronoun omission, and that some of those differences could in part be attributed to the type of contact situation in which they developed. However, other factors, such as their status as first or second languages, had to be taken into consideration as well.

Assuming that certain types of contact lead to simplification and that omitted pronouns result in simpler and more efficient structures, Chapter 4 aimed at exploring the distribution of pronoun deletion features in varieties of English by means of a cross-varietal study. The results showed a clear typological split between varieties of English as regards the attestation of pronoun deletion features: whereas primarily-L1 varieties (i.e., traditional low-contact and

dialect contact L1s) exhibited a low percentage of attestation of these grammatical features, L2-influenced varieties (i.e., shift L1s, indigenized L2s, pidgins, and creoles) displayed much higher attestation rates. More importantly, no further significant differences were found between the variety types included in each of these two groups of varieties. A geographical signal was also evident in the data, although only with respect to the attestation of some individual features and the pervasiveness of pronoun omission, that is, its frequency of occurrence in those varieties in which it is attested. In brief, Asian and Pacific varieties displayed a particularly strong inclination towards deletion vis-à-vis other world regions, which may have been caused by the influence of the substrate languages spoken in those areas. The findings of Chapter 4 suggest, however, that the most important explanatory factor for the attestation (or not) of pronoun omission features in varieties of English is whether they have a history of language contact. This in turn provides support for the claim that pronoun omission results in simplification without a loss in communicative efficiency, at least in structural terms. However, the fact remained that, in systemic terms, pronoun omission produces more complex grammars as it entails a larger set of referential expressions and, possibly, more rules to account for their use. Thus, omitted pronouns should be disfavoured by speakers of second-language varieties, as a more complex grammar may also require a greater learning effort, but this was clearly not the case in the cross-varietal study presented in Chapter 4. The findings of Chapter 5 neutralized this potential contradiction by showing that varieties which are mostly spoken as second languages displayed rather simple pronoun-deletion grammars, that is, the hypothesized increase in phenogrammatical system complexity did not actually take place in these varieties.

Chapter 5 addressed two issues that are crucial to understand the contribution of pronoun omission to the structural and system complexity of varieties of English. The first one was

related to the structural contexts in which speakers of (varieties of) English omit pronouns. If speakers produce pronouns in omitted form even in those contexts in which this may cause processing problems for hearers, then communication is no longer efficient, as any potential advantage for speakers derived from a reduction in articulatory effort is outweighed by the extra processing costs in which hearers may have to incur. Several environments in which omission would result in extra processing costs were identified in Chapter 5, and in all of these contexts overt pronouns were preferred, thus evidencing an efficient use of omission on the part of speakers. On the contrary, omitted pronouns were favoured when there were independent processing or cognitive considerations that facilitated omission and, more importantly, when they referred to highly accessible referents, as hearers may not have been able to coindex them with their correct antecedents if this was not the case.

In terms of the degree of structural complexity of the three varieties examined in Chapter 5, BrE was found to be more complex than either IndE or SgE, as omitted pronouns occurred more frequently in a wider range of contexts in the latter two varieties. SgE exhibited a higher overall frequency of omission than either IndE and BrE, and thus was the simplest of the three varieties from a structural perspective. Crucially, the spread of omitted pronouns to new contexts of use in the high-contact varieties did not occur in those environments in which omission would result in extra processing costs, which means that SgE and IndE became structurally simpler in an efficient manner.

The second issue addressed in Chapter 5 was related to the role that the language-external and language-internal constraints identified in Chapter 3 played in the varieties' pronoun-deletion grammars. Three random forests were computed, one per variety, to quantify the importance of the different factors that regulate the use of omitted and overt pronouns in English. The number of important constraints in each variety, that is, those that actually

discriminated between the omitted and overt variants, was taken to be an approximate measure of the varieties' degree of system complexity, with a larger number of significant factors indicating a more complex pronoun-deletion grammar. The results of the per variety random forests revealed that SgE was the most systemically complex of the varieties examined, followed by BrE. On the other hand, IndE, which was the only genuine second-language variety of the set, exhibited the simplest pronoun-deletion grammar. This cline of varieties as to their degree of system complexity clearly reflected the preferences of their users, since a more complex grammar is also in all likelihood harder to learn (and, possibly, to use too).

A great deal of grammatical homogeneity across varieties was evident in the results of the per variety random forests. The same three grammatical constraints, namely 'Coordination,' 'Clause position,' and 'Verb form,' were the ones with the strongest discriminatory power between omitted and overt pronouns in the three varieties, since they occupied the first three positions in the ranking. SgE and BrE exhibited highly comparable arrangements of constraints, as they assigned very similar relative importance values to the factors included in the analyses. IndE, on the other hand, deviated from the other two varieties as regards the relative importance of the less important constraints, thus displaying a more idiosyncratic pronoun-deletion grammar.

One of the most significant findings of the corpus-based study presented in Chapter 5 was that it uncovered a potential trade-off between structural and system complexity in SgE. On the one hand, SgE exhibited a higher overall frequency of pronoun omission and an extension of this feature to a wider range of contexts, two developments which represent simplification in structural terms. On the other hand, a large number of constraints was found to have a significant influence on the choice between omitted and overt pronouns, which complexifies

the pronoun-deletion grammar of SgE from a systemic perspective. One possible explanation for the existence of this complexity trade-off is that the increase in frequency of a structural variant within a set of alternatives and its extension to new contexts of use require the support of a more complex grammar with more rules determining when to use each of the members of the set. This increase in system complexity in turn entails a greater learning effort on the part of speakers, which may not be possible (or, at least, highly disfavoured) if they are second-language learners. This could be the reason why IndE differed from SgE in the overall frequency of pronoun deletion, despite the fact that both of them are high-contact varieties with omission-friendly substrate languages. In addition, as argued by Hawkins (2014, 87–89), different types of speakers may have different processing preferences which are then reflected in the grammars of their languages or dialects. Certain groups of speakers may exploit some grammatical resources that go unnoticed by other users and thus tilt the balance in favour of a particular linguistic variant in the case that, at a certain point in the history of a language or dialect, they become a sizeable group within the speech community. Second-language users should favour pronoun omission (in those contexts in which it enables a successful transmission of information), but if an increase in frequency of this structural variant requires the support of a larger number of grammatical rules or constraints, then any potential structural benefits may be cancelled out by the disadvantages of having to learn and use a more complex grammar. This is the situation in IndE, which is spoken as a second language by the majority of its users. However, if at a certain point the variety shifts from a second to a first language, as is arguably the case of SgE, the increase in system complexity may not be so problematic, thus paving the path to structural simplification. However, this explanation of the findings is yet a tentative one, as it is based on the study of only three varieties. To truly

verify or refute this hypothesis, the set of varieties examined needs to be expanded so as to include a wider range of variety types and sociolinguistic situations.

Finally, the results of Chapters 4 and 5 emphasized the importance of language contact in the distribution of omitted pronouns in varieties of English, but this important role of contact was argued to be mainly a reflection of the influence of second-language acquisition and use, and not of the substrate languages spoken alongside English in the relevant speech communities. That is, substrate effects were claimed to play a minor role in the frequency and distribution of omitted pronouns. This conclusion was supported, first, by the cross-varietal study of Chapter 4, which showed that region was not a very powerful predictor of the attestation of pronoun deletion features across 76 varieties of English; pronoun deletion attestation was instead better accounted for by variety type. Second, the corpus-based study of Chapter 5 revealed a high degree of grammatical homogeneity across the three varieties examined. On the one hand, only three interactions of ‘Variety’ and other predictors were found by the binary mixed-effects logistic regression, which means that most of the constraints included in the analysis behaved similarly in BrE, IndE, and SgE. These cross-varietally stable constraints were in turn those that were grounded in processing and cognitive motivations, which were not expected to vary as all users share the same cognitive architecture and thus the same memory and/or processing limitations. Likewise, the overall frequency of omission was similar across the varieties, especially between BrE and IndE. On the other hand, the varieties were also largely homogeneous with respect to the relative importance of the constraints included in the analysis. The rankings of predictors as to their relative importance were for the most part comparable in the three varieties, especially between BrE and SgE. Only IndE behaved somewhat differently in this respect.

Despite the convincing nature of these findings, we should not rush to the conclusion that substrate influence plays absolutely no role, or even just a minor one, in the occurrence of omitted pronouns in varieties of English. As argued by Sharma (2009, 190):

we cannot simply treat varieties that exhibit presence of a particular trait as identical, particularly where bilingual varieties of English are concerned. Surface similarities across New Englishes can be skin deep, diverging dramatically upon closer examination, due to substrate systems or substrate-superstrate interaction.

In fact, a more qualitative inspection of specific instances of pronoun omission, in line with Sharma's proposal, has already uncovered some patterns that seem to replicate those found in the substrate languages (cf. Schröter and Kortmann 2016). However, Mufwene (2008, 150; cf. Chapter 4, Section 4.1) cautions against establishing a direct link between correlation and causation since, as he argues, the mere presence of a grammatical feature in the substrate language(s) does not inevitably entail its transfer to the target language: many features are never transferred and others may occur at a certain point but never find their way into the communal norm. Other factors that account for the selection of some features over others in particular contact situations, such as complexity, frequency, saliency, or congruence between the substrate and the superstrate must necessarily be taken into consideration. A complete exploration of substrate effects lied beyond the scope of the present work, which aimed at characterizing pronoun omission in terms of its contribution to the structural and system complexity and efficiency of (varieties of) English. However, it is an issue of utmost importance to be able to disentangle the effects of simplification due to second-language acquisition and use from those of substrate influence. Future research may indeed focus on this particular topic by, for example, comparing the instances of omission in the varieties

examined with those found in the substrate languages to ascertain whether they are (dis)favoured in the relevant contexts. The (dis)preference for omission on the basis of what is found in the substrate languages could in turn be included as a variable in further multifactorial analyses to explore whether it truly affects the choice between omitted and overt pronouns and how it interacts with other predictors.

This monograph has contributed to the field of World Englishes by focusing on pronoun omission, a grammatical feature that had not yet been widely analysed in English. In addition, by approaching the study of pronoun omission from the domain of language complexity and communicative efficiency, the present work has examined this linguistic phenomenon from a perspective from which, to the best of my knowledge, it had not been considered before. Therefore, the results offered here can be said to fill a gap in the literature at the crossroads of research on varieties of English, complexity and efficiency studies, and pronoun omission. However, as is usually the case, the present analysis could be improved and expanded. Besides including a wider range of varieties, several steps could be taken to complement it with further work on the same topic. From a methodological perspective, the findings of the corpus-based study in Chapter 5 are based on a rather small sample of 2,456 relevant observations of omitted and overt pronouns (half of them were instances of omitted and the other half of overt pronouns). As explained in that chapter, the reason for this reduced dataset lay in the fact that omitted pronouns had to be identified by manually retrieving them from the corpus, which was a highly time-consuming task. Therefore, finding a way to automatically retrieve instances of omission from corpora would greatly ease the task of expanding the present dataset both in terms of varieties and observations. Furthermore, this would enable us to complement the instances obtained from *ICE* with those retrieved from larger corpora such as *GloWbE*. The structural contexts favouring pronoun omission that were identified in

Chapter 5 may be used to help us limit the search to a restricted set of environments, although we would possibly get many false positives. A larger sample of omitted and overt pronouns would permit the inclusion of more predictors and predictor interactions in the binary mixed-effects logistic regression. No model is ever perfect, despite the fact that ours was extremely sound and displayed an outstanding discriminating power; there is always a percentage of variation that is unaccounted for. I believe that by including interactions of ‘Variety,’ ‘Coordination’ or ‘Clause position,’ and the rest of the predictors already considered, the accuracy of the results would improve significantly, as they may uncover potential differences in the linguistic conditioning of cases in canonical contexts (i.e., in coordination and initial position) and in non-canonical ones (i.e. non-coordinate contexts and non-initial clause positions).

Referent accessibility, despite its prominence in previous research, turned out not to discriminate strongly between omitted and overt pronouns. Even though this result is not extremely surprising given that both referential expressions are markers of high accessibility, the question remains whether the way in which we operationalized accessibility, as a categorical variable with three levels (i.e., high, intermediate, and low), was not sensitive enough. A more fine-grained operationalization by, for instance, assigning an accessibility score to each antecedent based on the number of accessibility-increasing conditions at play, may serve to capture subtler differences between omitted and overt pronouns in this respect.

In addition, the assessment of the contribution of pronoun omission to the relative (i.e., processing) complexity and communicative efficiency of the grammar of (varieties of) English would greatly benefit from further support from experimental evidence. Psycholinguistic experiments could be set up to confirm the validity of the results achieved

here in a more controlled environment. They would also allow us to isolate the effects of the predictors to a greater extent, thus making our conclusions more accurate.

Finally, in line with the important role played by individual verb forms in the choice between omitted and overt pronouns, an interesting approach would be to account for pronoun omission in (varieties of) English in terms of some partially lexically filled constructions à la Torres Cacoullos and Travis (2014). This and the previous issues, however, cannot possibly be discussed here and must for the time being be left for future research.

REFERENCES AND SOURCES

REFERENCES

- Ariel, Mira. 1988. 'Referring and Accessibility'. *Journal of Linguistics* 24 (1): 65–87.
- . 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.
- . 1994. 'Interpreting Anaphoric Expression: A Cognitive versus a Pragmatic Approach'. *Journal of Linguistics* 30 (1): 3–42.
- . 2001. 'Accessibility Theory: An Overview'. In *Text Representation: Linguistic and Psycholinguistic Aspects*, edited by Ted J. M. Sanders, Joost Schliperoord, and Wilbert Spooren, 29–87. Amsterdam/Philadelphia: John Benjamins.
- Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge/New York: Cambridge University Press.
- Bao, Zhiming. 2001. 'The Origins of Empty Categories in Singapore English'. *Journal of Pidgin and Creole Languages* 16 (2): 275–319.
- . 2010. 'A Usage-Based Approach to Substratum Transfer: The Case of Four Unproductive Features in Singapore English'. *Language* 86 (4): 792–820.
- Barton, Kamil. 2018. *MuMIn: Multi-Model Inference* (version 1.40.4). <https://CRAN.R-project.org/package=MuMIn>.
- Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. 'Fitting Linear Mixed-Effects Models Using lme4'. *Journal of Statistical Software* 67 (1): 1–48.

- Bayley, Robert, and Otto Santa Ana. 2004. 'Chicano English: Morphology and Syntax'. In *A Handbook of Varieties of English*, edited by Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie, and Clive Upton, 374–90. Berlin/New York: De Gruyter Mouton.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Biewer, Carolin. 2015. *South Pacific Englishes: A Sociolinguistic and Morphosyntactic Profile of Fiji English, Samoan English and Cook Islands English*. Amsterdam/Philadelphia: John Benjamins.
- Bisang, Walter. 2009. 'On the Evolution of Complexity: Sometimes Less is More in East and Mainland Southeast Asia'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 34–49. Oxford/New York: Oxford University Press.
- Brunner, Thomas. 2014. 'Structural Nativization, Typology and Complexity: Noun Phrase Structures in British, Kenyan and Singaporean English'. *English Language and Linguistics* 18 (1): 23–48.
- . 2017. *Simplicity and Typological Effects in the Emergence of New Englishes: The Noun Phrase in Singaporean and Kenyan English*. Berlin/Boston: De Gruyter Mouton.
- Butterworth, Guy, and Evelyn M. Hatch. 1978. 'A Spanish Speaking Adolescent's Acquisition of English Syntax'. In *Second Language Acquisition: A Book of Readings*, edited by Evelyn M. Hatch, 231–45. Rowley (MA): Newbury House.

- Bybee, Joan L. 1998. 'The Emergent Lexicon'. In *CLS 34: The Panels*, edited by M. Catherine Gruber, Kenneth S. Olson, and Derrick Higgins, 421–35. Chicago: Chicago Linguistic Society.
- . 2001. 'Frequency Effects on French Liaison'. In *Frequency and the Emergence of Linguistic Structure*, edited by Joan L. Bybee and Paul J. Hopper, 337–59. Amsterdam/Philadelphia: John Benjamins.
- . 2002. 'Word Frequency and Context of Use in the Lexical Diffusion of Phonetically Conditioned Sound Change'. *Language Variation and Change* 14 (3): 261–90.
- . 2006. 'From Usage to Grammar: The Mind's Response to Repetition'. *Language* 82 (4): 711–33.
- . 2007. *Frequency of Use and the Organization of Language*. Oxford/New York: Oxford University Press.
- Bybee, Joan L., and Joanne Scheibman. 1999. 'The Effect of Usage on Degrees of Constituency: The Reduction of *Don't* in American English'. *Linguistics* 37 (4): 575–96.
- Bybee, Joan L., and Sandra A. Thompson. 1997. 'Three Frequency Effects in Syntax'. In *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society, February 14–17, 1997: General Session and Parasession on Pragmatics and Grammatical Structure*, edited by Matthew L. Juge and Jeri L. Moxley, 378–88. Berkeley: Berkeley Linguistics Society.
- Chipere, Ngoni. 2003. *Understanding Complex Sentences: Native Speaker Variation in Syntactic Competence*. Basingstoke: Palgrave Macmillan.

- . 2009. 'Individual Differences in Processing Complex Grammatical Structures'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 178–91. Oxford/New York: Oxford University Press.
- Chomsky, Noam. 1980. *Rules and Representations*. Oxford: Blackwell.
- . 1981. *Lectures on Government and Binding*. Dordrecht/Providence: Foris Publications.
- . 1982. *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge (MA): The MIT Press.
- . 1991. 'Linguistics and Cognitive Science: Problems and Mysteries'. In *The Chomskyan Turn*, edited by Asa Kasher, 26–53. Oxford: Blackwell.
- Claes, Jeroen. 2017. 'Probabilistic Grammar: The View from Cognitive Sociolinguistics'. *Glossa: A Journal of General Linguistics* 2 (1): 1–30.
- Cole, Melvyn D. 2009. 'Null Subjects: A Reanalysis of the Data'. *Linguistics* 47 (3): 559–87.
- . 2010. 'Thematic Null Subjects and Accessibility'. *Studia Linguistica* 64 (3): 271–320.
- Cysouw, Michael. 2013. 'Disentangling Geography from Genealogy'. In *Space in Language and Linguistics: Geographical, Interactional and Cognitive Perspectives*, edited by Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi 21–37. Berlin/Boston: De Gruyter.
- Dąbrowska, Ewa. 1997. 'The LAD Goes to School: A Cautionary Tale for Nativists'. *Linguistics* 35 (4): 735–66.
- Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam/Philadelphia: John Benjamins.

- Davydova, Julia. 2011. *The Present Perfect in Non-Native Englishes: A Corpus-Based Study of Variation*. Berlin/Boston: De Gruyter Mouton.
- Deutscher, Guy. 2000. *Syntactic Change in Akkadian: The Evolution of Sentential Complementation*. Oxford: Oxford University Press.
- . 2009. “‘Overall Complexity’: A Wild Goose Chase?” In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 243–51. Oxford/New York: Oxford University Press.
- Dittmar, Norbert. 1984. ‘Semantic Features of Pidginized Learner Varieties of German’. In *Second Languages: A Cross-Linguistic Perspective*, edited by Roger W. Andersen, 243–70. Rowley (MA): Newbury House.
- Dušková, Libuše. 1969. ‘On Sources of Errors in Foreign Language Learning’. *International Review of Applied Linguistics* 7 (1): 11–36.
- Ellis, Nick C. 1998. ‘Emergentism, Connectionism and Language Learning.’ *Language Learning* 48 (4): 631–64.
- Fakhri, Ahmed. 1984. ‘The Use of Communicative Strategies in Narrative Discourse: A Case Study of a Learner of Moroccan Arabic as a Second Language’. *Language Learning* 34 (3): 15–37.
- Felix, Sascha W. 1980. ‘Interference, Interlanguage and Related Issues’. In *Second Language Development: Trends and Issues*, edited by Sascha W. Felix, 93–110. Tübingen: Gunter Narr Verlag.
- Fenk-Oczlon, Gertraud, and August Fenk. 2008. ‘Complexity Trade-Offs between the Subsystems of Language’. In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, 43–66. Amsterdam/Philadelphia: John Benjamins.

- Filipović, Luna, and John A. Hawkins. 2013. 'Multiple Factors in Second Language Acquisition: The CASP Model'. *Linguistics* 51 (1): 145–76.
- Frascarelli, Mara. 2007. 'Subjects, Topics and the Interpretation of Referential *pro*: An Interface Approach to the Linking of (Null) Pronouns'. *Natural Language & Linguistic Theory* 25 (4): 691–734.
- Fuchs, Robert. 2016. 'The Frequency of the Present Perfect in Varieties of English around the World'. In *Re-assessing the Present Perfect: Corpus Studies and Beyond*, edited by Valentin Werner, Elena Seoane, and Cristina Suárez-Gómez, 223–58. Berlin/Boston: De Gruyter.
- Gil, David. 2008. 'How Complex Are Isolating Languages?' In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, 109–32. Amsterdam/Philadelphia: John Benjamins.
- . 2009. 'How Much Grammar Does It Take to Sail a Boat?' In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 19–33. Oxford/New York: Oxford University Press.
- Givón, Talmy. 1983. 'Topic Continuity in Discourse: An Introduction'. In *Topic Continuity in Discourse: An Introduction*, edited by Talmy Givón, 5–41. Amsterdam/Philadelphia: John Benjamins.
- Gries, Stefan Th. 2005. 'Syntactic Priming: A Corpus-Based Approach'. *Journal of Psycholinguistic Research* 34 (4): 365–99.
- . 2007. *Coll.Analysis 3.2a: A Program for R for Windows 2.X*.
- . 2015. 'The Most Under-Used Method in Corpus Linguistics: Multi-Level (and Mixed-Effects) Models'. *Corpora* 10 (1): 95–125.

- . 2016. 'Variationist Analysis: Variability Due to Random Effects and Autocorrelation'. In *Triangulating Methodological Approaches in Corpus Linguistic Research*, edited by Paul Baker and Jesse A. Egbert. New York: Routledge.
- Gundel, Jeanette K., Nancy Stenson, and Elaine Tarone. 1984. 'Acquiring Pronouns in a Second Language: Evidence for Hypothesis Testing'. *Studies in Second Language Acquisition* 6 (2): 215–25.
- Gundel, Jeanette K., and Elaine Tarone. 1983. 'Language Transfer and the Acquisition of Pronominal Anaphora'. In *Language Transfer and Language Learning*, edited by Susan M. Gass and Larry Selinker, 281–86. Rowley (MA): Newbury House.
- Haegeman, Liliane. 1990. 'Understood Subjects in English Diaries: On the Relevance of Theoretical Syntax for the Study of Register Variation'. *Multilingua* 9 (2): 157–99.
- . 1995. *Introduction to Government and Binding Theory*. 2nd ed. Oxford: Blackwell Publishers.
- Haiman, John. 1994. 'Ritualization and the Development of Language'. In *Perspectives on Grammaticalization*, edited by William Pagliuca, 3–24. Amsterdam: John Benjamins.
- Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Cambridge/New York: Cambridge University Press.
- . 1999. 'Processing Complexity and Filler-Gap Dependencies across Grammars'. *Language* 75 (2): 244–85.
- . 2004. *Efficiency and Complexity in Grammars*. Oxford/New York: Oxford University Press.
- . 2009. 'An Efficiency Theory of Complexity and Related Phenomena'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 252–68. Oxford/New York: Oxford University Press.

- . 2014. *Cross-Linguistic Variation and Efficiency*. Oxford/New York: Oxford University Press.
- Heine, Bernd, and Tania Kuteva. 2005. *Language Contact and Grammatical Change*. Cambridge/New York: Cambridge University Press.
- Heller, Benedikt, Benedikt Szmrecsanyi, and Jason Grafmiller. 2017. 'Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation Across Varieties of English'. *Journal of English Linguistics* 45 (1): 3–27.
- Hockett, Charles F. 1958. *A Course in Modern Linguistics*. New York: Macmillan.
- Holmberg, Anders, Aarti Nayudu, and Michelle Sheehan. 2009. 'Three Partial Null-Subject Languages: A Comparison of Brazilian Portuguese, Finnish and Marathi'. *Studia Linguistica* 63 (1): 59–97.
- Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. 2006. 'Survival Ensembles'. *Biostatistics* 7 (3): 355–73.
- Huang, C.-T. James. 1984. 'On the Distribution and Reference of Empty Pronouns'. *Linguistic Inquiry* 15 (4): 531–74.
- Huang, Yan. 1992. 'Against Chomsky's Typology of Empty Categories'. *Journal of Pragmatics* 17 (1): 1–29.
- . 2000. *Anaphora: A Cross-Linguistic Approach*. Oxford/New York: Oxford University Press.
- Huber, Magnus. 2012. 'Syntactic and Variational Complexity in British and Ghanaian English: Relative Clause Formation in the Written Parts of the International Corpus of English'. In *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, edited by Bernd Kortmann and Benedikt Szmrecsanyi, 218–42. Berlin/Boston: De Gruyter.

- Huddleston, Rodney, Geoffrey K. Pullum, Laurie Bauer, Betty Birner, Ted Briscoe, Peter Collins, David Denison, David Lee, Anita Mittwoch, Geoffrey Nunberg, Frank Palmer, John Payne, Peter Peterson, Lesley Stirling, and Gregory Ward. 2002. *The Cambridge Grammar of the English Language*. Cambridge/New York: Cambridge University Press.
- Huebner, Thorn. 1983. 'System and Variability in Interlanguage Syntax'. *Language Learning* 35 (2): 141–63.
- Jaeger, T. Florian, and Harry Tily. 2011. 'On Language "Utility": Processing Complexity and Communicative Efficiency'. *Wiley Interdisciplinary Reviews: Cognitive Science* 2 (3): 323–35.
- Jaeggli, Osvaldo, and Kenneth J. Safir. 1989. 'Subject Parameter and Parametric Theory'. In *The Null Subject Parameter*, edited by Osvaldo Jaeggli and Kenneth J. Safir. Dordrecht/Boston/London: Kluwer Academic Publishers.
- Kachru, Yamuna. 2006. *Hindi*. Amsterdam/Philadelphia: John Benjamins.
- Karlsson, Fred, Matti Miestamo, and Kaius Sinnemäki. 2008. 'Introduction: The Problem of Language Complexity'. In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, VII–XIV. Amsterdam/Philadelphia: John Benjamins.
- Klein, Wolfgang, and Clive Perdue. 1997. 'The Basic Variety (or: Couldn't Natural Languages Be Much Simpler?)'. *Second Language Research* 13 (4): 301–47.
- Kortmann, Bernd. 2006. 'Syntactic Variation in English: A Global Perspective'. In *The Handbook of English Linguistics*, edited by Bas Aarts and April McMahon, 603–24. Oxford: Blackwell Publishers.

- . 2013. 'How Powerful Is Geography as an Explanatory Factor in Morphosyntactic Variation? Areal Features in the Anglophone World'. In *Space in Language and Linguistics: Geographical, Interactional and Cognitive Perspectives*, 165–94. Berlin/Boston: De Gruyter.
- Kortmann, Bernd, and Kerstin Lunkenheimer. 2012. 'Introduction'. In *The Mouton World Atlas of Variation in English*, edited by Bernd Kortmann and Kerstin Lunkenheimer, 1–11. Berlin/Boston: De Gruyter.
- Kortmann, Bernd, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie, and Clive Upton, eds. 2004. *A Handbook of Varieties of English*. Berlin/New York: De Gruyter Mouton.
- Kortmann, Bernd, and Benedikt Szmrecsanyi. 2009. 'World Englishes between Simplification and Complexification'. In *World Englishes Problems, Properties and Prospects: Selected Papers from the 13th IAWC Conference*, edited by Lucia Siebers and Thomas Hoffmann, 265–85. Amsterdam/Philadelphia: John Benjamins.
- . 2011. 'Parameters of Morphosyntactic Variation in World Englishes: Prospects and Limitations of Searching for Universals'. In *Linguistic Universals and Language Variation*, edited by Peter Siemund, 264–90. Berlin/Boston: De Gruyter Mouton.
- Krug, Manfred. 1998. 'String Frequency: A Cognitive Motivating Factor in Coalescence, Language Processing and Linguistic Change'. *Journal of English Linguistics* 26 (4): 286–320.
- Kusters, Christiaan Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.
- . 2008. 'Complexity in Linguistic Theory, Language Learning and Language Change'. In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo,

- Kaius Sinnemäki, and Fred Karlsson, 3–22. Amsterdam/Philadelphia: John Benjamins.
- Larsen-Freeman, Diane. 1997. 'Chaos/Complexity Science and Second Language Acquisition'. *Applied Linguistics* 18 (2): 141–65.
- Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, Variation, and Usage*. Cambridge/New York: Cambridge University Press.
- Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge (MA): MIT Press.
- Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam/Philadelphia: John Benjamins.
- Li, Charles N., and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley/Los Angeles/London: University of California Press.
- Maas, Utz. 2009. 'Orality versus Literacy as a Dimension of Complexity'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 164–77. Oxford/New York: Oxford University Press.
- MacWhinney, Brian. 2005. 'A Unified Model of Language Acquisition'. In *Handbook of Bilingualism: Psycholinguistic Approaches*, edited by Judith F. Kroll and Annette M. B. de Groot, 49–67. Oxford/New York: Oxford University Press.
- McWhorter, John H. 2001. 'The World's Simplest Grammars Are Creole Grammars'. *Linguistic Typology* 5 (2): 125–66.
- . 2007. *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars*. Oxford/New York: Oxford University Press.

- . 2008. 'Why Does a Language Undress? Strange Cases in Indonesia'. In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, 167–90. Amsterdam/Philadelphia: John Benjamins.
- Mesthrie, Rajend, and Rakesh M. Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge/New York: Cambridge University Press.
- Miestamo, Matti. 2006a. 'On the Complexity of Standard Negation'. In *A Man of Measure: Festschrift in Honour of Fred Karlsson on His 60th Birthday*, edited by Mickael Suominen, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen, and Kaius Sinnemäki, 345–56. Turku: The Linguistic Association of Finland.
- . 2006b. 'On the Feasibility of Complexity Metrics'. In *FinEst Linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics. Tallinn, May 6–7, 2004*, edited by Krista Kerge and Maria-Maren Sepper, 11–26. Tallinn: Tallinn University Press.
- . 2008. 'Grammatical Complexity in a Cross-Linguistic Perspective'. In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, 23–41. Amsterdam/Philadelphia: John Benjamins.
- Miestamo, Matti, Kaius Sinnemäki, and Fred Karlsson, eds. 2008. *Language Complexity: Typology, Contact, Change*. Amsterdam/Philadelphia: John Benjamins.
- Moag, Rodney F., and Robert Poletto. 1991. 'Discourse Level Evidence for South Asia as a Linguistic Area'. *Language Sciences* 13 (2): 229–54.
- Mondorf, Britta. 2009. *More Support for More-Support: The Role of Processing Constraints on the Choice between Synthetic and Analytic Comparative Forms*. Amsterdam/Philadelphia: John Benjamins.

- Mufwene, Salikoko S. 2008. *Language Evolution: Contact, Competition and Change*. Continuum.
- Mukherjee, Joybrato. 2007. 'Steady States in the Evolution of New Englishes: Present-Day Indian English as an Equilibrium'. *Journal of English Linguistics* 35 (2): 157–87.
- Nagy, Naomi G, Nina Aghdasi, Derek Denis, and Alexandra Motut. 2011. 'Null Subjects in Heritage Languages: Contact Effects in a Cross-Linguistic Context'. *University of Pennsylvania Working Papers in Linguistics* 17 (2): 135–44.
- Neeleman, Ad, and Kriszta Szendrői. 2007. 'Radical Pro Drop and the Morphology of Pronouns'. *Linguistic Inquiry* 38 (4): 671–714.
- Newmeyer, Frederick J., and Laurel B. Preston. 2014. 'Introduction'. In *Measuring Grammatical Complexity*, edited by Frederick J. Newmeyer and Laurel B. Preston, First edition, 1–13. Oxford/New York: Oxford University Press.
- Nichols, Johanna. 2009. 'Linguistic Complexity: A Comprehensive Definition and Survey'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 110–25. Oxford/New York: Oxford University Press.
- Nihalani, Paroo, Ray Tongue, and Priya Hosali. 1977. *Indian and British English*. Delhi: Oxford University Press.
- Otheguy, Ricardo, Ana Celia Zentella, and David Livert. 2007. 'Language and Dialect Contact in Spanish in New York: Toward the Formation of a Speech Community'. *Language* 83 (4): 770–802.
- Piantadosi, Steven T., Harry Tily, and Edward Gibson. 2012. 'The Communicative Function of Ambiguity in Language'. *Cognition* 122 (3): 280–91.

- Posio, Pekka. 2014. 'Subject Expression in Grammaticalizing Constructions: The Case of *Creo* and *Acho* "I Think" in Spanish and Portuguese'. *Journal of Pragmatics* 63: 5–18.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London/New York: Longman.
- R Core Team. 2017. *R: A Language and Environment for Statistical Computing* (version 3.4.3). Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Radford, Andrew. 2004. *Minimalist Syntax: Exploring the Structure of English*. Cambridge: Cambridge University Press.
- Real Academia Española. 2009. *Nueva Gramática de la Lengua Española. Vol. 1: Morfología y Sintaxis*. Madrid: Espasa Libros.
- Rescher, Nicholas. 1998. *Complexity: A Philosophical Overview*. New Brunswick/London: Transaction Publishers.
- Rizzi, Luigi. 1986. 'Null Objects in Italian and the Theory of *pro*'. *Linguistic Inquiry* 17 (3): 501–57.
- Rohdenburg, Günter. 1996. 'Cognitive Complexity and Increased Grammatical Explicitness in English'. *Cognitive Linguistics* 7 (2): 149–82.
- Röthlisberger, Melanie, Jason Grafmiller, and Benedikt Szendrői. 2017. 'Cognitive Indigenization Effects in the English Dative Alternation'. *Cognitive Linguistics* 28 (4).
- Ruppenhofer, Josef, and Laura A. Michaelis. 2010. 'A Constructional Account of Genre-Based Argument Omissions'. *Constructions and Frames* 2 (2): 158–84.
- Rusten, Kristian A. 2013. 'Empty Referential Subjects in Old English Prose: A Quantitative Analysis'. *English Studies* 94 (8): 970–92.

- . 2014. 'Null Referential Subjects from Old to Early Modern English'. In *'Ye Whom the Charms of Grammar Please': Studies in English Language History in Honour of Leiv Egil Breivik*, edited by Kari E. Haugland, Kevin McCafferty, and Kristian A. Rusten, 249–70. Oxford: Peter Lang.
- . 2015. 'A Quantitative Study of Empty Referential Subjects in Old English Prose and Poetry'. *Transactions of the Philological Society* 113 (1): 53–75.
- Sampson, Geoffrey. 2001. 'Demographic Correlates of Complexity in British Speech'. In *Empirical Linguistics*, edited by Geoffrey Sampson, 57–73. London: Continuum.
- . 2009. 'A Linguistic Axiom Challenged'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 1–18. Oxford/New York: Oxford University Press.
- Sampson, Geoffrey, David Gil, and Peter Trudgill, eds. 2009. *Language Complexity as an Evolving Variable*. Oxford/New York: Oxford University Press.
- Sato, Yosuke, and Chonghyuck Kim. 2012. 'Radical Pro Drop and the Role of Syntactic Agreement in Colloquial Singapore English'. *Lingua* 122 (8): 858–73.
- Scheibman, Joanne. 2000. 'I Dunno: A Usage-Based Account of the Phonological Reduction of *Don't* in American English Conversation'. *Journal of Pragmatics* 32 (1): 105–24.
- Schneider, Agnes. 2012. 'Typological Profile: Pidgins and Creoles'. In *The Mouton Atlas of Variation in English*, edited by Bernd Kortmann and Kerstin Lunkenheimer, 875–904. Berlin/Boston: De Gruyter.
- Schneider, Edgar W. 2003. 'The Dynamics of New Englishes: From Identity Construction to Dialect Birth'. *Language* 79 (2): 233–81.
- . 2007. *Postcolonial English: Varieties around the World*. Cambridge/New York: Cambridge University Press.

- . 2013. 'English as a Contact Language: The "New Englishes"'. In *English as a Contact Language*, edited by Daniel Schreier and Marianne Hundt, 131–48. Cambridge: Cambridge University Press.
- Schröter, Verena, and Bernd Kortmann. 2016. 'Pronoun Deletion in Hong Kong English and Colloquial Singaporean English'. *World Englishes* 35 (2): 221–41.
- Schumann, John. 1984. 'Non-Syntactic Speech in Spanish-English Basilang'. In *Second Languages: A Cross-Linguistic Perspective*, edited by Roger W. Andersen, 355–74. Rowley (MA): Newbury House.
- Scott, Mike. 2012. *WordSmith Tools* (version 6). Liverpool: Lexical Analysis Software.
- Selinger, Jessica C., Shawn M. O'Connor, Jeremy D. Wong, and J. Maxwell Donelan. 2015. 'Humans Can Continuously Optimize Energetic Cost during Walking'. *Current Biology* 25 (18): 2452–56.
- Sharma, Devyani. 2009. 'Typological Diversity in New Englishes'. *English World-Wide* 30 (2): 170–95.
- . 2012. 'Indian English'. In *The Mouton World Atlas of Variation in English*, edited by Bernd Kortmann and Kerstin Lunkenheimer, 523–30. Berlin/Boston: De Gruyter.
- Shin, Naomi L. 2014. 'Grammatical Complexification in Spanish in New York: 3sg Pronoun Expression and Verbal Ambiguity'. *Language Variation and Change* 26 (3): 303–30.
- Silva-Corvalán, Carmen. 1997. 'Variación Sintáctica en el Discurso Oral: Problemas Metodológicos'. In *Trabajos de Sociolingüística Hispánica*, edited by Francisco Moreno Fernández, 115–35. Alcalá de Henares: Universidad de Alcalá.
- . 2001. *Sociolingüística y Pragmática del Español*. Washington DC: Georgetown University Press.

- Silveira, Agripino S. 2007. 'Patterns of Subjectivity in Spoken Brazilian Portuguese'. In *Ensaaios Provindos do Corpus Oral do Português de Fortaleza*, edited by Clemilton Lopes Pinheiro and Kilpatric M. R. Campelo. Fortaleza: Eduece.
- . 2008. 'Frequency Effects, Specialization of Forms, and Subject Expression in Brazilian Portuguese'. In *Proceedings of the VII High Desert Linguistics Society Conference*, edited by Evan Ashworth and Daniel Sanford. Albuquerque: Department of Linguistics, University of New Mexico.
- Sinnemäki, Kaius. 2008. 'Complexity Trade-Offs in Core Argument Marking'. In *Language Complexity: Typology, Contact, Change*, edited by Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, 67–88. Amsterdam/Philadelphia: John Benjamins.
- Slobin, Dan I. 1977. 'Language in Childhood and in History'. In *Language Learning and Thought*, edited by John Macnamara, 185–214. New York: Academic Press.
- Speas, Margaret. 1994. 'Null Arguments in a Theory of Economy of Projection'. In *University of Massachusetts Occasional Papers in Linguistics 17*, edited by Elena Benedicto and Jeff Runner. Amherst: Graduate Linguistics Student Association.
- . 2006. 'Economy, Agreement, and the Representation of Null Arguments'. In *Arguments and Agreement*, edited by Peter Ackema, Patrick Brandt, Maaike Schoorlemmer, and Fred Weerman, 35–75. Oxford/New York: Oxford University Press.
- Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. 'Conditional Variable Importance for Random Forests'. *BMC Bioinformatics* 9 (307).

- Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. 'Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution'. *BMC Bioinformatics* 8 (25).
- Suárez-Gómez, Cristina. 2017. 'Transparency and Language Contact in the Nativization of Relative Clauses in New Englishes'. *English World-Wide* 38 (2): 211–37.
- . 2017. 'The Coding of Perfect Meaning in African, Asian and Caribbean Englishes'. *English Language and Linguistics*: 1–21.
- Szmrecsanyi, Benedikt. 2005. 'Language Users as Creatures of Habit: A Corpus-Based Analysis of Persistence in Spoken English'. *Corpus Linguistics and Linguistic Theory* 1 (1): 113–49.
- . 2006. *Morphosyntactic Persistence in Spoken English. A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin/New York: De Gruyter Mouton.
- . 2009. 'Typological Parameters of Intralingual Variability: Grammatical Analyticity versus Syntheticity in Varieties of English'. *Language Variation and Change* 21 (3): 319–53.
- Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller, and Melanie Röthlisberger. 2016. 'Around the World in Three Alternations: Modeling Syntactic Variation in Varieties of English'. *English World-Wide* 37 (2): 109–37.
- Szmrecsanyi, Benedikt, and Bernd Kortmann. 2009a. 'Between Simplification and Complexification: Non-Standard Varieties of English around the World'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 64–79. Oxford/New York: Oxford University Press.

- . 2009b. 'The Morphosyntax of Varieties of English Worldwide: A Quantitative Perspective'. *Lingua* 119 (11): 1643–63.
- . 2012. 'Introduction: Linguistic Complexity - Second Language Acquisition, Indigenization, Contact'. In *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, edited by Bernd Kortmann and Benedikt Szmrecsanyi, 6–34. Berlin/Boston: De Gruyter.
- Tagliamonte, Sali A., and R. Harald Baayen. 2012. 'Models, Forests, and Trees of York English: Was/Were Variation as a Case Study for Statistical Practice'. *Language Variation and Change* 24 (2): 135–78.
- Tily, Harry, and Steven Piantadosi. 2009. 'Refer Efficiently: Use Less Informative Expressions for More Predictable Meanings'. In *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference*, edited by Kees van Deemter, Albert Gatt, Roger van Gompel, and Emiel Krahmer, 1–8. Amsterdam: Cognitive Science Society.
- Tomioka, Satoshi. 2003. 'The Semantics of Japanese Null Pronouns and Its Cross-Linguistic Implications'. In *The Interfaces: Deriving and Interpreting Omitted Structures*, edited by Kerstin Schwabe and Susanne Winkler. Amsterdam/Philadelphia: John Benjamins.
- Toole, Janine. 1996. 'The Effect of Genre on Referential Choice'. In *Reference and Referent Accessibility*, edited by Thornstein Fretheim and Jeanette K. Gundel, 263–90. Amsterdam/Philadelphia: John Benjamins.
- Torres Cacoullos, Rena, and Catherine E. Travis. 2014. 'Prosody, Priming and Particular Constructions: The Patterning of English First-Person Singular Subject Expression in Conversation'. *Journal of Pragmatics* 63: 19–34.

- Travis, Catherine E., and Rena Torres Cacoullos. 2012. 'What Do Subject Pronouns Do in Discourse? Cognitive, Mechanical and Constructional Factors in Variation'. *Cognitive Linguistics* 23 (4): 711–48.
- Trudgill, Peter. 2009a. 'Sociolinguistic Typology and Complexification'. In *Language Complexity as an Evolving Variable*, edited by Geoffrey Sampson, David Gil, and Peter Trudgill, 98–109. Oxford/New York: Oxford University Press.
- . 2009b. 'Vernacular Universals and the Sociolinguistic Typology of English Dialects'. In *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, edited by Markku Filppula, Juhani Klemola, and Heli Paulasto, 304–22. London: Routledge.
- . 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.
- Von Humboldt, Wilhelm. 1836. *Über Die Verschiedenheit Des Menschlichen Sprachbaues Und Ihren Einfluss Auf Die Geistige Entwicklung Des Menschengeschlechts*. Berlin: Dümmler.
- Wagner, Susanne. 2018. 'Never Saw One – First-person Null Subjects in Spoken English', *English Language and Linguistics* 22 (1): 1–34.
- Walkden, George. 2013. 'Null Subjects in Old English'. *Language Variation and Change* 25 (2): 155–78.
- Walkden, George, and Kristian A. Rusten. 2017. 'Null Subjects in Middle English'. *English Language and Linguistics* 21 (3): 439–73.
- Wasow, Thomas, Amy Perfors, and David Beaver. 2005. 'The Puzzle of Ambiguity'. In *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe*,

- edited by C. Orhan Orgun and Peter Sells, 265–82. California: Center for the Study of Language and Information Publications.
- Wee, Lionel. 2004. 'Singapore English: Morphology and Syntax'. In *A Handbook of Varieties of English*, edited by Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie, and Clive Upton, 1058–72. Berlin/New York: De Gruyter Mouton.
- Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag New York.
- Williams, Jessica. 1987. 'Non-Native Varieties of English: A Special Case of Language Acquisition'. *English World-Wide* 8 (2): 161–99.
- . 1988. 'Zero Anaphora in Second Language Acquisition: A Comparison among Three Varieties of English'. *Studies in Second Language Acquisition* 10 (3): 339–70.
- . 1989. 'Pronoun Copies, Pronominal Anaphora and Zero Anaphora in Second Language Production'. In *Variation in Second Language Acquisition*, edited by Susan M. Gass, Carolyn Madden, Dennis Preston, and Larry Selinker, 153–89. Clevedon/Philadelphia: Multilingual Matters.
- Xu, Liejiong. 1986. 'Free Empty Category'. *Linguistic Inquiry* 17 (1): 75–93.
- Xu, Liejiong, and D. Terence Langendoen. 1985. 'Topic Structures in Chinese'. *Language* 61 (1): 1–27.
- Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.
- Zobl, Helmut. 1984. 'Uniformity and Source-Language Variation across Developmental Continua'. In *Language Universals and Second Language Acquisition*, edited by William E. Rutherford, 185–218. Amsterdam/Philadelphia: John Benjamins.

- Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.
- Zyzik, Eve C. 2008. 'Null Objects in Second Language Acquisition: Grammatical vs. Performance Models'. *Second Language Research* 24 (1): 65–110.

SOURCES

- Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries (GloWbE)*. <https://corpus.byu.edu/glowbe/>. (Accessed on 11-04-2018)
- Dryer, Matthew S. 2013. 'Expression of Pronominal Subjects'. In *The World Atlas of Language Structures Online*, edited by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/101>. (Accessed on 11-04-2018)
- Dryer, Matthew S., and Martin Haspelmath, eds. 2013. *The World Atlas of Language Structures Online*. Leipzig: Oxford University Press. <http://wals.info>. (Accessed on 11-04-2018)
- ICE-GB = *International Corpus of English - the British Component*. 1990. Project Coordinated by Prof. Gerald Nelson at the Chinese University of Hong Kong. <http://www.icecorpora.net/ice/download.htm>
- ICE-IND = *International Corpus of English - the Indian Component*. 2002. Project Coordinated by Prof. S. V. Shastri at Shivaji University and Prof. Dr. Gerhard Leitner at Freie Universität Berlin. <http://www.ice-corpora.net/ice/download.htm>
- ICE-SIN = *International Corpus of English – the Singaporean Component*. 2002. Project Coordinated by Prof. Paroo Nihilani, Dr. Ni Yibin, Dr. Anne Pakir and Dr Vincent

Ooi. Singapore: The National University of Singapore. <http://ice-corpora.net/ice/download.htm>

Kortmann, Bernd, and Kerstin Lunkenheimer, eds. 2013. *The Electronic World Atlas of Varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ewave-atlas.org/>. (Accessed on 11-04-2018)

Michaelis, Susanne M., Philippe Maurer, Martin Haspelmath, and Magnus Huber, eds. 2013. *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://apics-online.info>. (Accessed on 11-04-2018)

APPENDIX

Table 31. Model summary: variety type as predictor of pronoun omission attestation

<i>Model summary</i>	
C-index of concordance	0.741
Correct predictions	76.32%
Likelihood ratio χ^2	13.09
Degrees of freedom	5
<i>p</i> -value	< 0.05

Table 32. Contrasts: variety type as predictor of pronoun omission attestation

<i>Reference value = Traditional L1</i>				
Predictor	Estimate	Std. error	Z	<i>p</i>
Intercept	−0.406	0.646	−0.628	0.530
Type = Dialect-contact L1	−0.105	0.975	−0.108	0.914
Type = Shift L1	2.197	1.000	2.197	< 0.05
Type = Indigenized L2	2.485	0.990	2.511	< 0.05
Type = Pidgin	1.322	1.057	1.251	0.211
Type = Creole	1.435	0.830	1.730	0.084
<i>Reference value = Dialect-contact L1</i>				
Predictor	Estimate	Std. error	Z	<i>p</i>

Intercept	−0.511	0.730	−0.699	0.484
Type = Traditional L1	0.105	0.975	0.108	0.914
Type = Shift L1	2.303	1.057	2.179	< 0.05
Type = Indigenized L2	2.590	1.047	2.474	< 0.05
Type = Pidgin	1.427	1.111	1.285	0.199
Type = Creole	1.540	0.897	1.717	0.086
Reference value = Shift L1				
Predictor	Estimate	Std. error	Z	p
Intercept	1.792	0.764	2.346	< 0.05
Type = Traditional L1	−2.197	1.000	−2.197	< 0.05
Type = Dialect-contact L1	−2.303	1.057	−2.179	< 0.05
Type = Indigenized L2	0.288	1.070	0.269	0.788
Type = Pidgin	−0.876	1.133	−0.773	0.440
Type = Creole	−0.762	0.925	−0.824	0.410
Reference value = Indigenized L2				
Predictor	Estimate	Std. error	Z	p
Intercept	2.079	0.750	2.773	< 0.01
Type = Traditional L1	−2.485	0.990	−2.511	< 0.05
Type = Dialect-contact L1	−2.590	1.047	−2.474	< 0.05
Type = Shift L1	−0.288	1.070	−0.269	0.788
Type = Pidgin	−1.163	1.124	−1.035	0.301
Type = Creole	−1.050	0.913	−1.150	0.250
Reference value = Pidgin				
Predictor	Estimate	Std. error	Z	p
Intercept	0.916	0.837	1.095	0.273
Type = Traditional L1	−1.322	1.057	−1.251	0.211

Type = Dialect–contact L1	–1.427	1.111	–1.285	0.199
Type = Shift L1	0.876	1.133	0.773	0.440
Type = Indigenized L2	1.163	1.124	1.035	0.301
Type = Creole	0.113	0.986	0.115	0.908

Table 33. Model summary: region as predictor of pronoun omission pervasiveness

<i>Model summary</i>	
Multiple R^2	0.445
Adjusted R^2	0.361
Residual standard error	0.953
Residual degrees of freedom	46
F	5.23
Degrees of freedom	7
p -value	< 0.001

Table 34. Contrasts: region as predictor of pronoun omission pervasiveness

<i>Reference value = British Isles</i>				
Predictor	Estimate	Std. error	t	p
Intercept	1.120	0.426	2.628	< 0.05
Region = America	–0.234	0.558	–0.420	0.677
Region = Africa	–0.362	0.507	–0.713	0.479
Region = Asia	1.780	0.543	3.277	< 0.01
Region = Australia	0.480	0.639	0.751	0.457
Region = Pacific	1.497	0.577	2.594	< 0.05
Region = Caribbean	0.300	0.522	0.575	0.568

Region = South Atlantic	0.880	0.797	1.104	0.275
Reference value = America				
Predictor	Estimate	Std. error	t	p
Intercept	0.886	0.360	2.459	< 0.05
Region = British Isles	0.234	0.558	0.420	0.677
Region = Africa	-0.127	0.453	-0.281	0.780
Region = Asia	2.014	0.493	4.084	< 0.001
Region = Australia	0.714	0.597	1.196	0.238
Region = Pacific	1.731	0.530	3.265	< 0.01
Region = Caribbean	0.534	0.470	1.138	0.261
Region = South Atlantic	1.114	0.764	1.458	0.152
Reference value = Africa				
Predictor	Estimate	Std. error	t	p
Intercept	0.758	0.275	2.757	< 0.01
Region = British Isles	0.362	0.507	0.713	0.479
Region = America	0.127	0.453	0.281	0.780
Region = Asia	2.142	0.435	4.924	< 0.001
Region = Australia	0.842	0.550	1.530	0.133
Region = Pacific	1.858	0.477	3.900	< 0.001
Region = Caribbean	0.662	0.408	1.622	0.112
Region = South Atlantic	1.242	0.728	1.706	0.095
Reference value = Asia				
Predictor	Estimate	Std. error	t	p
Intercept	2.900	0.337	8.608	< 0.001
Region = British Isles	-1.780	0.543	-3.277	< 0.01
Region = America	-2.014	0.493	-4.084	< 0.001

Region = Africa	−2.142	0.435	−4.924	< 0.001
Region = Australia	−1.300	0.584	−2.228	< 0.05
Region = Pacific	−0.283	0.515	−0.551	0.585
Region = Caribbean	−1.480	0.452	−3.274	< 0.01
Region = South Atlantic	−0.900	0.753	−1.195	0.238
Reference value = Australia				
Predictor	Estimate	Std. error	t	p
Intercept	1.600	0.477	3.358	< 0.01
Region = British Isles	−0.480	0.639	−0.751	0.457
Region = America	−0.714	0.597	−1.196	0.238
Region = Africa	−0.842	0.550	−1.530	0.133
Region = Asia	1.300	0.584	2.228	< 0.05
Region = Pacific	1.017	0.615	1.653	0.105
Region = Caribbean	−0.180	0.564	−0.319	0.751
Region = South Atlantic	0.400	0.825	0.485	0.630
Reference value = Pacific				
Predictor	Estimate	Std. error	t	p
Intercept	2.61	0.389	6.726	< 0.001
Region = British Isles	−1.49	0.577	−2.594	< 0.05
Region = America	−1.731	0.530	−3.265	< 0.01
Region = Africa	−1.858	0.477	−3.900	< 0.001
Region = Asia	0.283	0.515	0.551	0.585
Region = Australia	−1.017	0.615	−1.653	0.105
Region = Caribbean	−1.197	0.492	−2.432	< 0.05
Region = South Atlantic	−0.617	0.778	−0.793	0.432
Reference value = Caribbean				

Predictor	Estimate	Std. error	<i>t</i>	<i>p</i>
Intercept	1.420	0.301	4.712	< 0.001
Region = British Isles	−0.300	0.522	−0.575	0.568
Region = America	−0.534	0.470	−1.138	0.261
Region = Africa	−0.662	0.408	−1.622	0.112
Region = Asia	1.480	0.452	3.274	< 0.01
Region = Australia	0.180	0.564	0.319	0.751
Region = Pacific	1.197	0.492	2.432	< 0.05
Region = South Atlantic	0.580	0.738	0.786	0.436

Table 35. Model summary: region as predictor of object omission attestation

<i>Model summary</i>	
C-index of concordance	0.838
Correct predictions	78.95%
Likelihood ratio χ^2	32.90
Degrees of freedom	7
<i>p</i> -value	< 0.001

Table 36. Contrasts: region as predictor of object omission attestation

<i>Reference value = British Isles</i>				
Predictor	Estimate	Std. error	<i>Z</i>	<i>p</i>
Intercept	−2.398	1.045	−2.296	< 0.05
Region = America	0.201	1.484	0.135	0.892
Region = Africa	2.041	1.155	1.767	0.077
Region = Asia	19.964	1398.721	0.014	0.989

Region = Australia	1.992	1.387	1.436	0.151
Region = Pacific	3.497	1.326	2.637	< 0.01
Region = Caribbean	0.693	1.297	0.534	0.593
Region = South Atlantic	1.705	1.610	1.059	0.290
Reference value = America				
Predictor	Estimate	Std. error	Z	p
Intercept	−2.197	1.054	−2.084	< 0.05
Region = British Isles	−0.201	1.484	−0.135	0.892
Region = Africa	1.841	1.164	1.582	0.114
Region = Asia	19.763	1398.721	0.014	0.989
Region = Australia	1.792	1.394	1.285	0.199
Region = Pacific	3.296	1.333	2.472	< 0.05
Region = Caribbean	0.493	1.305	0.377	0.706
Region = South Atlantic	1.504	1.616	0.931	0.352
Reference value = Africa				
Predictor	Estimate	Std. error	Z	p
Intercept	−0.357	0.493	−0.724	0.469
Region = British Isles	−2.041	1.155	−1.767	0.077
Region = America	−1.841	1.164	−1.582	0.114
Region = Asia	17.922	1398.721	0.013	0.990
Region = Australia	−0.049	1.037	−0.047	0.963
Region = Pacific	1.455	0.954	1.526	0.127
Region = Caribbean	−1.348	0.913	−1.476	0.140
Region = South Atlantic	−0.336	1.320	−0.255	0.799
Reference value = Asia				
Predictor	Estimate	Std. error	Z	p

Intercept	17.570	1398.720	0.013	0.990
Region = British Isles	−19.960	1398.720	−0.014	0.989
Region = America	−19.760	1398.720	−0.014	0.989
Region = Africa	−17.920	1398.720	−0.013	0.990
Region = Australia	−17.970	1398.720	−0.013	0.990
Region = Pacific	−16.470	1398.720	−0.012	0.991
Region = Caribbean	−19.270	1398.720	−0.014	0.989
Region = South Atlantic	−18.260	1398.720	−0.013	0.990
Reference value = Australia				
Predictor	Estimate	Std. error	Z	p
Intercept	−0.405	0.913	−0.444	0.657
Region = British Isles	−1.992	1.387	−1.436	0.151
Region = America	−1.792	1.394	−1.285	0.199
Region = Africa	0.049	1.037	0.047	0.962
Region = Asia	17.972	1398.721	0.013	0.990
Region = Pacific	1.504	1.225	1.228	0.219
Region = Caribbean	−1.299	1.193	−1.089	0.276
Region = South Atlantic	−0.288	1.528	−0.188	0.851
Reference value = Pacific				
Predictor	Estimate	Std. error	Z	p
Intercept	1.099	0.817	1.346	0.178
Region = British Isles	−3.497	1.326	−2.637	< 0.01
Region = America	−3.296	1.333	−2.472	< 0.05
Region = Africa	−1.455	0.954	−1.526	0.127
Region = Asia	16.468	1398.721	0.012	0.991
Region = Australia	−1.504	1.225	−1.228	0.219

Region = Caribbean	−2.803	1.121	−2.500	< 0.05
Region = South Atlantic	−1.792	1.472	−1.217	0.224
Reference value = Caribbean				
Predictor	Estimate	Std. error	Z	p
Intercept	−1.705	0.769	−2.218	< 0.05
Region = British Isles	−0.693	1.297	−0.534	0.593
Region = America	−0.493	1.305	−0.377	0.706
Region = Africa	1.348	0.913	1.476	0.140
Region = Asia	19.271	1398.721	0.014	0.989
Region = Australia	1.299	1.193	1.089	0.276
Region = Pacific	2.803	1.121	2.500	< 0.05
Region = South Atlantic	1.012	1.446	0.700	0.484

Table 37. Model summary: region as predictor of object omission pervasiveness

Model summary	
Multiple R^2	0.476
Adjusted R^2	0.293
Residual standard error	0.228
Residual degrees of freedom	20
F	2.599
Degrees of freedom	7
p -value	< 0.05

Table 38. Contrasts: region as predictor of object omission pervasiveness

Reference value = British Isles

Predictor	Estimate	Std. error	<i>t</i>	<i>p</i>
Intercept	1.000	0.228	4.381	< 0.001
Region = America	−0.700	0.323	−2.169	< 0.05
Region = Africa	−0.514	0.244	−2.108	< 0.05
Region = Asia	−0.200	0.242	−0.826	0.418
Region = Australia	−0.200	0.280	−0.715	0.483
Region = Pacific	−0.267	0.247	−1.082	0.292
Region = Caribbean	−0.700	0.280	−2.504	< 0.05
Region = South Atlantic	−0.400	0.323	−1.239	0.230
<i>Reference value = America</i>				
Predictor	Estimate	Std. error	<i>t</i>	<i>p</i>
Intercept	0.300	0.228	1.314	0.204
Region = British Isles	0.700	0.323	2.169	< 0.05
Region = Africa	0.186	0.244	0.761	0.456
Region = Asia	0.500	0.242	2.065	0.052
Region = Australia	0.500	0.280	1.789	0.089
Region = Pacific	0.433	0.247	1.758	0.094
Region = Caribbean	−2.78e−16	0.280	0.000	1.000
Region = South Atlantic	0.300	0.323	0.929	0.364
<i>Reference value = Africa</i>				
Predictor	Estimate	Std. error	<i>t</i>	<i>p</i>
Intercept	0.486	0.086	5.630	< 0.001
Region = British Isles	0.514	0.244	2.108	< 0.05
Region = America	−0.186	0.244	−0.761	0.456
Region = Asia	0.314	0.118	2.661	< 0.05
Region = Australia	0.314	0.183	1.717	0.101

Region = Pacific	0.248	0.127	1.950	0.065
Region = Caribbean	−0.186	0.183	−1.015	0.322
Region = South Atlantic	0.114	0.244	0.468	0.645
Reference value = Asia				
Predictor	Estimate	Std. error	t	p
Intercept	0.800	0.081	9.914	< 0.001
Region = British Isles	0.200	0.242	0.826	0.419
Region = America	−0.500	0.242	−2.065	0.052
Region = Africa	−0.314	0.118	−2.661	< 0.05
Region = Australia	9.86e−17	0.180	0.000	1.000
Region = Pacific	−0.067	0.123	−0.541	0.595
Region = Caribbean	−0.500	0.180	−2.771	< 0.05
Region = South Atlantic	−0.200	0.242	−0.826	0.419
Reference value = Australia				
Predictor	Estimate	Std. error	t	p
Intercept	0.800	0.161	4.957	< 0.001
Region = British Isles	0.200	0.280	0.715	0.483
Region = America	−0.500	0.280	−1.789	0.089
Region = Africa	−0.314	0.183	−1.717	0.101
Region = Asia	−1.87e−16	0.180	0.000	1.000
Region = Pacific	−0.067	0.186	−0.358	0.724
Region = Caribbean	−0.500	0.228	−2.191	< 0.05
Region = South Atlantic	−0.200	0.280	−0.715	0.483
Reference value = Pacific				
Predictor	Estimate	Std. error	t	p
Intercept	0.733	0.093	7.870	< 0.001

Region = British Isles	0.267	0.247	1.082	0.292
Region = America	−0.433	0.247	−1.758	0.094
Region = Africa	−0.248	0.127	−1.950	0.065
Region = Asia	0.067	0.123	0.541	0.595
Region = Australia	0.067	0.186	0.358	0.724
Region = Caribbean	−0.433	0.186	−2.325	< 0.05
Region = South Atlantic	−0.133	0.247	−0.541	0.595
Reference value = Caribbean				
Predictor	Estimate	Std. error	t	p
Intercept	0.300	0.161	1.859	0.078
Region = British Isles	0.700	0.280	2.504	< 0.05
Region = America	−1.20e−16	0.280	0.000	1.000
Region = Africa	0.186	0.183	1.015	0.322
Region = Asia	0.500	0.180	2.771	< 0.05
Region = Australia	0.500	0.228	2.191	< 0.05
Region = Pacific	0.433	0.186	2.325	< 0.05
Region = South Atlantic	0.300	0.280	1.073	0.296

Table 39. Model summary: region as predictor of non-referential subject omission attestation

Model summary	
C-index of concordance	0.781
Correct predictions	77.63%
Likelihood ratio χ^2	21.7
Degrees of freedom	7
p-value	< 0.01

Table 40. Contrasts: region as predictor of non-referential subject omission attestation

<i>Reference value = British Isles</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−1.609	0.775	−2.078	< 0.05
Region = America	0.762	1.037	0.735	0.463
Region = Africa	0.069	1.002	0.069	0.945
Region = Asia	3.555	1.320	2.693	< 0.01
Region = Australia	2.015	1.197	1.683	0.092
Region = Pacific	2.120	1.064	1.992	< 0.05
Region = Caribbean	−0.095	1.091	−0.087	0.930
Region = South Atlantic	2.303	1.449	1.589	0.112
<i>Reference value = America</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−0.847	0.690	−1.228	0.220
Region = British Isles	−0.762	1.037	−0.735	0.463
Region = Africa	−0.693	0.939	−0.738	0.460
Region = Asia	2.793	1.272	2.195	< 0.05
Region = Australia	1.253	1.144	1.095	0.274
Region = Pacific	1.358	1.005	1.352	0.177
Region = Caribbean	−0.858	1.033	−0.830	0.407
Region = South Atlantic	1.540	1.406	1.096	0.273
<i>Reference value = Africa</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−1.540	0.636	−2.421	< 0.05
Region = British Isles	−0.069	1.002	−0.069	0.945
Region = America	0.693	0.939	0.738	0.460

Region = Asia	3.486	1.244	2.802	< 0.01
Region = Australia	1.946	1.112	1.749	0.080
Region = Pacific	2.051	0.969	2.118	< 0.05
Region = Caribbean	−0.164	0.998	−0.165	0.869
Region = South Atlantic	2.233	1.380	1.618	0.105
Reference value = Asia				
Predictor	Estimate	Std. error	Z	p
Intercept	1.946	1.069	1.820	0.069
Region = British Isles	−3.555	1.320	−2.693	< 0.01
Region = America	−2.793	1.272	−2.195	< 0.05
Region = Africa	−3.486	1.244	−2.802	< 0.01
Region = Australia	−1.540	1.406	−1.096	0.273
Region = Pacific	−1.435	1.295	−1.108	0.268
Region = Caribbean	−3.651	1.317	−2.773	< 0.01
Region = South Atlantic	−1.253	1.626	−0.771	0.441
Reference value = Australia				
Predictor	Estimate	Std. error	Z	p
Intercept	0.406	0.913	0.444	0.657
Region = British Isles	−2.015	1.197	−1.683	0.092
Region = America	−1.253	1.144	−1.095	0.274
Region = Africa	−1.946	1.113	−1.749	0.080
Region = Asia	1.540	1.406	1.096	0.273
Region = Pacific	0.105	1.169	0.090	0.928
Region = Caribbean	−2.110	1.193	−1.768	0.077
Region = South Atlantic	0.288	1.528	0.188	0.851
Reference value = Pacific				

Predictor	Estimate	Std. error	Z	p
Intercept	0.511	0.730	0.699	0.484
Region = British Isles	-2.120	1.065	-1.992	< 0.05
Region = America	-1.358	1.005	-1.352	0.177
Region = Africa	-2.051	0.969	-2.118	< 0.05
Region = Asia	1.435	1.295	1.108	0.268
Region = Australia	-0.105	1.169	-0.090	0.928
Region = Caribbean	-2.216	1.060	-2.090	< 0.05
Region = South Atlantic	0.182	1.426	0.128	0.898
Reference value = Caribbean				
Predictor	Estimate	Std. error	Z	p
Intercept	-1.705	0.769	-2.218	< 0.05
Region = British Isles	0.095	1.091	0.087	0.930
Region = America	0.857	1.033	0.830	0.407
Region = Africa	0.164	0.998	0.165	0.869
Region = Asia	3.651	1.317	2.773	< 0.01
Region = Australia	2.110	1.193	1.768	0.077
Region = Pacific	2.216	1.060	2.090	< 0.05
Region = South Atlantic	2.398	1.446	1.658	0.097

Table 41. Model summary: region as predictor of referential *it* omission attestation

Model summary	
C-index of concordance	0.782
Correct predictions	71.05%
Likelihood ratio χ^2	20.98

Degrees of freedom	7
<i>p</i> -value	< 0.01

Table 42. Contrasts: region as predictor of referential *it* omission attestation

Reference value = British Isles				
Predictor	Estimate	Std. error	Z	<i>p</i>
Intercept	−2.398	1.044	−2.296	< 0.05
Region = America	1.992	1.228	1.623	0.105
Region = Africa	1.522	1.172	1.299	0.194
Region = Asia	2.398	1.261	1.901	0.057
Region = Australia	−15.168	1769.258	−0.009	0.993
Region = Pacific	2.398	1.261	1.901	0.057
Region = Caribbean	3.602	1.235	2.917	< 0.05
Region = South Atlantic	3.091	1.610	1.920	0.055
Reference value = America				
Predictor	Estimate	Std. error	Z	<i>p</i>
Intercept	−0.406	0.646	−0.628	0.530
Region = British Isles	−1.992	1.228	−1.623	0.105
Region = Africa	−0.470	0.837	−0.562	0.574
Region = Asia	0.406	0.957	0.423	0.672
Region = Australia	−17.161	1769.258	−0.010	0.992
Region = Pacific	0.406	0.957	0.423	0.672
Region = Caribbean	1.609	0.922	1.746	0.081
Region = South Atlantic	1.099	1.384	0.794	0.428
Reference value = Africa				

Predictor	Estimate	Std. error	Z	p
Intercept	−0.876	0.532	−1.645	0.100
Region = British Isles	−1.522	1.172	−1.299	0.194
Region = America	0.470	0.837	0.562	0.574
Region = Asia	0.876	0.885	0.989	0.323
Region = Australia	−16.691	1769.258	−0.009	0.992
Region = Pacific	0.876	0.885	0.989	0.323
Region = Caribbean	2.079	0.847	2.456	< 0.05
Region = South Atlantic	1.569	1.335	1.175	0.240
Reference value = Asia				
Predictor	Estimate	Std. error	Z	p
Intercept	−2.39e−16	0.707	0.000	1.000
Region = British Isles	−2.398	1.261	−1.901	0.057
Region = America	−0.406	0.957	−0.423	0.672
Region = Africa	−0.876	0.885	−0.989	0.323
Region = Australia	−17.570	1769	−0.010	0.992
Region = Pacific	1.86e−16	1.000	0.000	1.000
Region = Caribbean	1.204	0.966	1.246	0.213
Region = South Atlantic	0.693	1.414	0.490	0.624
Reference value = Australia				
Predictor	Estimate	Std. error	Z	p
Intercept	−17.570	1769.260	−0.010	0.992
Region = British Isles	15.170	1769.260	0.009	0.993
Region = America	17.160	1769.260	0.010	0.992
Region = Africa	16.690	1769.260	0.009	0.992
Region = Asia	17.570	1769.260	0.010	0.992

Region = Pacific	17.570	1769.260	0.010	0.992
Region = Caribbean	18.770	1769.260	0.011	0.992
Region = South Atlantic	18.260	1769.260	0.010	0.992
Reference value = Pacific				
Predictor	Estimate	Std. error	Z	p
Intercept	7.77e-16	0.707	0.000	1.000
Region = British Isles	-2.398	1.261	-1.901	0.057
Region = America	-0.406	0.957	-0.423	0.672
Region = Africa	-0.876	0.885	-0.989	0.323
Region = Asia	-7.79e-16	1.000	0.000	1.000
Region = Australia	-17.570	1769	-0.010	0.992
Region = Caribbean	1.204	0.966	1.246	0.213
Region = South Atlantic	0.693	1.414	0.490	0.624
Reference value = Caribbean				
Predictor	Estimate	Std. error	Z	p
Intercept	1.204	0.658	1.829	0.067
Region = British Isles	-3.602	1.235	-2.917	< 0.01
Region = America	-1.609	0.922	-1.746	0.081
Region = Africa	-2.079	0.847	-2.456	< 0.05
Region = Asia	-1.204	0.966	-1.246	0.213
Region = Australia	-18.770	1769.2578	-0.011	0.992
Region = Pacific	-1.204	0.966	-1.246	0.213
Region = South Atlantic	-0.511	1.390	-0.367	0.713

Table 43. Model summary: region as predictor of non-referential *it* omission attestation

Model summary	
C-index of concordance	0.774
Correct predictions	76.32%
Likelihood ratio χ^2	16.11
Degrees of freedom	7
<i>p</i> -value	< 0.05

Table 44. Contrasts: region as predictor of non-referential *it* omission attestation

Reference value = British Isles				
Predictor	Estimate	Std. error	Z	p
Intercept	−2.398	1.045	−2.296	< 0.05
Region = America	0.201	1.484	0.135	0.892
Region = Africa	1.219	1.191	1.024	0.306
Region = Asia	2.909	1.275	2.282	< 0.05
Region = Australia	−15.168	1769.2579	−0.009	0.993
Region = Pacific	2.398	1.261	1.901	0.057
Region = Caribbean	1.587	1.205	1.317	0.188
Region = South Atlantic	3.091	1.610	1.920	0.055
Reference value = America				
Predictor	Estimate	Std. error	Z	p
Intercept	−2.197	1.054	−2.084	< 0.05
Region = British Isles	−0.201	1.484	−0.135	0.892
Region = Africa	1.017	1.199	0.849	0.396
Region = Asia	2.708	1.282	2.112	< 0.05

Region = Australia	−15.369	1769.2579	−0.009	0.993
Region = Pacific	2.197	1.269	1.731	0.083
Region = Caribbean	1.386	1.213	1.143	0.253
Region = South Atlantic	2.890	1.616	1.789	0.074
Reference value = Africa				
Predictor	Estimate	Std. error	Z	p
Intercept	−1.179	0.572	−2.061	< 0.05
Region = British Isles	−1.219	1.191	−1.024	0.306
Region = America	−1.019	1.199	−0.849	0.396
Region = Asia	1.690	0.928	1.822	0.069
Region = Australia	−16.387	1769.2577	−0.009	0.993
Region = Pacific	1.179	0.909	1.296	0.195
Region = Caribbean	0.368	0.830	0.443	0.658
Region = South Atlantic	1.872	1.352	1.385	0.166
Reference value = Asia				
Predictor	Estimate	Std. error	Z	p
Intercept	0.511	0.730	0.699	0.484
Region = British Isles	−2.909	1.275	−2.282	< 0.05
Region = America	−2.708	1.282	−2.112	< 0.05
Region = Africa	−1.690	0.928	−1.822	0.069
Region = Australia	−18.077	1769.2578	−0.010	0.992
Region = Pacific	−0.511	1.017	−0.503	0.615
Region = Caribbean	−1.322	0.946	−1.398	0.162
Region = South Atlantic	0.182	1.426	0.128	0.898
Reference value = Australia				
Predictor	Estimate	Std. error	Z	p

Intercept	−17.570	1769.260	−0.010	0.992
Region = British Isles	15.170	1769.260	0.009	0.993
Region = America	15.370	1769.260	0.009	0.993
Region = Africa	16.390	1769.260	0.009	0.993
Region = Asia	18.080	1769.260	0.010	0.992
Region = Pacific	17.570	1769.260	0.010	0.992
Region = Caribbean	16.760	1769.260	0.009	0.992
Region = South Atlantic	18.260	1769.260	0.010	0.992
Reference value = Pacific				
Predictor	Estimate	Std. error	Z	p
Intercept	−3.18e−17	0.707	0.000	1.000
Region = British Isles	−2.394	1.261	−1.901	0.057
Region = America	−2.197	1.269	−1.731	0.083
Region = Africa	−1.179	0.909	−1.296	0.195
Region = Asia	0.511	1.017	0.503	0.615
Region = Australia	−17.570	1769	−0.010	0.992
Region = Caribbean	−0.811	0.928	−0.874	0.382
Region = South Atlantic	0.693	1.414	0.490	0.624
Reference value = Caribbean				
Predictor	Estimate	Std. error	Z	p
Intercept	−0.811	0.601	−1.349	0.177
Region = British Isles	−1.587	1.205	−1.317	0.188
Region = America	−1.386	1.213	−1.143	0.253
Region = Africa	−0.368	0.830	−0.443	0.658
Region = Asia	1.322	0.946	1.398	0.162
Region = Australia	−16.755	1769.2577	−0.009	0.992

Region = Pacific	0.811	0.928	0.874	0.382
Region = South Atlantic	1.504	1.364	1.103	0.270

Table 45. Model summary: region as predictor of non-referential *it* omission pervasiveness

<i>Model summary</i>	
Multiple R^2	0.583
Adjusted R^2	0.404
Residual standard error	0.224
Residual degrees of freedom	14
F	3.263
Degrees of freedom	6
p -value	< 0.05

Table 46. Contrasts: region as predictor of non-referential *it* omission pervasiveness

<i>Reference value = British Isles</i>				
Predictor	Estimate	Std. error	t	p
Intercept	0.300	0.224	1.341	0.201
Region = America	1.03e-15	0.316	0.000	1.000
Region = Africa	3.10e-16	0.250	0.000	1.000
Region = Asia	0.380	0.245	1.551	0.143
Region = Pacific	0.500	0.250	1.999	0.065
Region = Caribbean	0.525	0.250	2.099	0.054
Region = South Atlantic	0.150	0.274	0.548	0.593
<i>Reference value = America</i>				
Predictor	Estimate	Std. error	t	p

Intercept	0.300	0.224	1.341	0.201
Region = British Isles	−6.26e−16	0.316	0.000	1.000
Region = Africa	−8.53e−16	0.250	0.000	1.000
Region = Asia	0.380	0.245	1.551	0.143
Region = Pacific	0.500	0.250	1.999	0.065
Region = Caribbean	0.525	0.250	2.099	0.054
Region = South Atlantic	0.150	0.274	0.548	0.593
Reference value = Africa				
Predictor	Estimate	Std. error	t	p
Intercept	0.300	0.112	2.682	< 0.05
Region = British Isles	2.76e−16	0.250	0.000	1.000
Region = America	1.25e−15	0.250	0.000	1.000
Region = Asia	0.380	0.150	2.532	< 0.05
Region = Pacific	0.500	0.158	3.161	< 0.01
Region = Caribbean	0.525	0.158	3.319	< 0.01
Region = South Atlantic	0.150	0.194	0.774	0.452
Reference value = Asia				
Predictor	Estimate	Std. error	t	p
Intercept	0.680	0.100	6.798	< 0.001
Region = British Isles	−0.380	0.245	−1.551	0.143
Region = America	−0.380	0.245	−1.551	0.143
Region = Africa	−0.380	0.150	−2.532	< 0.05
Region = Pacific	0.120	0.150	0.800	0.437
Region = Caribbean	0.145	0.150	0.966	0.350
Region = South Atlantic	−0.230	0.187	−1.229	0.239
Reference value = Pacific				

Predictor	Estimate	Std. error	<i>t</i>	<i>p</i>
Intercept	0.800	0.112	7.153	< 0.001
Region = British Isles	−0.500	0.250	−1.999	0.065
Region = America	−0.500	0.250	−1.999	0.065
Region = Africa	−0.500	0.158	−3.161	< 0.01
Region = Asia	−0.120	0.150	−0.800	0.437
Region = Caribbean	0.025	0.158	0.158	0.877
Region = South Atlantic	−0.350	0.194	−1.807	0.092
<i>Reference value = Caribbean</i>				
Predictor	Estimate	Std. error	<i>t</i>	<i>p</i>
Intercept	0.825	0.112	7.376	< 0.001
Region = British Isles	−0.525	0.250	−2.099	0.054
Region = America	−0.525	0.250	−2.099	0.054
Region = Africa	−0.525	0.158	−3.319	< 0.01
Region = Asia	−0.145	0.150	−0.966	0.350
Region = Pacific	−0.025	0.158	−0.158	0.877
Region = South Atlantic	−0.375	0.194	−1.936	0.073

Table 47. Further ‘Variety’ contrasts

<i>Reference value: IndE</i>				
Predictor	Estimate	Std. error	<i>Z</i>	<i>p</i>
Intercept	−2.464	0.388	−6.347	< 0.001
Variety = BrE	−1.775	0.644	−2.755	< 0.01
Variety = SgE	1.228	0.460	2.669	< 0.01
Text type = Spoken formal	0.108	0.432	0.249	0.803

Text type = Written informal	−0.633	0.472	−1.341	0.180
Text type = Written formal	−0.499	0.668	−0.746	0.456
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.177	0.444	11.670	< 0.001
Clause position = Initial	2.183	0.330	6.611	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. for.	−0.848	0.671	−1.264	0.206
Var. = SgE : Text type = Sp. for.	−0.742	0.603	−1.230	0.219
Var. = BrE : Text type = Wr. inf.	0.845	0.628	1.345	0.179
Var. = SgE : Text type = Wr. inf.	1.628	0.583	2.795	< 0.01
Var. = BrE : Text type = Wr. for.	0.539	0.930	0.580	0.562
Var. = SgE : Text type = Wr. for.	0.035	0.902	0.039	0.969
Var. = BrE : Coord. = Coord.	0.565	0.606	0.932	0.351
Var. = SgE : Coord. = Coord.	−1.091	0.579	−1.909	0.056
Var. = BrE : Cl. pos. = Initial	1.544	0.581	2.658	< 0.01
Var. = SgE : Cl. pos. = Initial	−0.661	0.399	−1.658	0.097
Reference value: SgE				
Predictor	Estimate	Std. error	Z	p
Intercept	−1.236	0.316	−3.907	< 0.001
Variety = BrE	−3.002	0.610	−4.920	< 0.001

Variety = IndE	−1.228	0.460	−2.669	< 0.01
Text type = Spoken formal	−0.634	0.421	−1.506	0.132
Text type = Written informal	0.995	0.352	2.825	< 0.01
Text type = Written formal	−0.463	0.625	−0.741	0.459
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	4.086	0.417	9.797	< 0.001
Clause position = Initial	1.528	0.237	6.425	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. for.	−0.106	0.665	−0.159	0.874
Var. = IndE : Text type = Sp. for.	0.742	0.603	1.231	0.218
Var. = BrE : Text type = Wr. inf.	−0.784	0.540	−1.451	0.147
Var. = IndE : Text type = Wr. inf.	−1.628	0.582	−2.795	< 0.01
Var. = BrE : Text type = Wr. for.	0.504	0.899	0.560	0.575
Var. = IndE : Text type = Wr. for.	−0.035	0.902	−0.039	0.969
Var. = BrE : Coord. = Coord.	1.656	0.603	2.747	< 0.01
Var. = IndE : Coord. = Coord.	1.092	0.572	1.909	0.056
Var. = BrE : Cl. pos. = Initial	2.205	0.534	4.127	< 0.001
Var. = IndE : Cl. pos. = Initial	0.662	0.399	1.658	0.097

Table 48. Further ‘Text type’ contrasts

<i>Reference value: BrE, Spoken Formal</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−4.979	0.671	−7.422	< 0.001
Variety = IndE	2.623	0.756	3.471	< 0.001
Variety = SgE	3.108	0.750	4.147	< 0.001
Text type = Spoken informal	0.740	0.516	1.435	0.151
Text type = Written informal	0.951	0.528	1.801	0.072
Text type = Written formal	0.781	0.713	1.095	0.274
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	0.031
Coordination = Coordination	5.742	0.483	11.890	< 0.001
Clause position = Initial	3.727	0.487	7.655	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.368	< 0.001
Priming = NP/No priming	0.218	0.165	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = IndE : Text type = Sp. inf.	−0.848	0.671	−1.264	0.206
Var. = SgE : Text type = Sp. inf.	−0.106	0.665	−0.159	0.873
Var. = IndE : Text type = Wr. inf.	−1.692	0.726	−2.330	< 0.05
Var. = SgE : Text type = Wr. inf.	0.678	0.676	1.003	0.316
Var. = IndE : Text type = Wr. for.	−1.387	0.985	−1.409	0.159
Var. = SgE : Text type = Wr. for.	−0.610	0.965	−0.632	0.527
Var. = IndE : Coord. = Coord.	−0.565	0.606	−0.932	0.351

Var. = SgE : Coord. = Coord.	−1.656	0.603	−2.747	< 0.01
Var. = IndE : Cl. pos. = Initial	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Initial	−2.205	0.534	−4.128	< 0.001
Reference value: BrE, Written Informal				
Predictor	Estimate	Std. error	Z	p
Intercept	−4.028	0.575	−7.005	< 0.001
Variety = IndE	0.930	0.728	1.278	0.201
Variety = SgE	3.786	0.638	5.939	< 0.001
Text type = Spoken informal	−0.211	0.419	−0.503	0.615
Text type = Spoken formal	−0.951	0.528	−1.801	0.072
Text type = Written formal	−0.170	0.662	−0.257	0.797
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	0.031
Coordination = Coordination	5.742	0.483	11.890	< 0.001
Clause position = Initial	3.727	0.487	7.655	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.368	< 0.001
Priming = NP/No priming	0.218	0.165	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = IndE : Text type = Sp. inf.	0.845	0.628	1.345	0.179
Var. = SgE : Text type = Sp. inf.	−0.784	0.540	−1.451	0.147
Var. = IndE : Text type = Sp. for.	1.692	0.727	2.329	< 0.05
Var. = SgE : Text type = Sp. for.	−0.678	0.676	−1.002	0.316
Var. = IndE : Text type = Wr. for.	0.305	0.939	0.325	0.745

Var. = SgE : Text type = Wr. for.	−1.288	0.898	−1.434	0.151
Var. = IndE : Coord. = Coord.	−0.565	0.606	−0.932	0.351
Var. = SgE : Coord. = Coord.	−1.656	0.603	−2.747	< 0.01
Var. = IndE : Cl. pos. = Initial	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Initial	−2.205	0.534	−4.128	< 0.001
Reference values: IndE, Spoken formal				
Predictor	Estimate	Std. error	Z	p
Intercept	−2.356	0.425	−5.545	< 0.001
Variety = BrE	−2.623	0.756	−3.471	< 0.001
Variety = SgE	0.486	0.547	0.889	0.374
Text type = Spoken informal	−0.108	0.431	−0.249	0.803
Text type = Written informal	−0.741	0.508	−1.459	0.144
Text type = Written formal	−0.606	0.684	−0.886	0.376
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.177	0.444	11.670	< 0.001
Clause position = Initial	2.183	0.330	6.611	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp.inf.	0.848	0.671	1.264	0.206
Var. = SgE : Text type = Sp. inf.	0.742	0.603	1.231	0.218
Var. = BrE : Text type = Wr. inf.	1.692	0.726	2.330	< 0.05

Var. = SgE : Text type = Wr. inf.	2.370	0.668	3.550	< 0.001
Var. = BrE : Text type = Wr. for.	1.387	0.985	1.408	0.159
Var. = SgE : Text type = Wr. for.	0.777	0.945	0.822	0.411
Var. = BrE : Coord. = Coord.	0.565	0.606	0.932	0.351
Var. = SgE : Coord. = Coord.	−1.091	0.579	−1.909	0.056
Var. = BrE : Cl. pos. = Initial	1.544	0.581	2.658	< 0.01
Var. = SgE : Cl. pos. = Initial	−0.661	0.399	−1.658	0.097
Reference value: SgE, Spoken formal				
Predictor	Estimate	Std. error	Z	p
Intercept	−1.87	0.409	−4.569	< 0.001
Variety = BrE	−3.108	0.750	−4.146	< 0.001
Variety = IndE	−0.486	0.547	−0.888	0.374
Text type = Spoken informal	0.634	0.421	1.506	0.132
Text type = Written informal	1.629	0.431	3.778	< 0.001
Text type = Written formal	0.171	0.656	0.260	0.795
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	4.086	0.417	9.797	< 0.001
Clause position = Initial	1.528	0.237	6.425	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. inf.	0.106	0.665	0.159	0.873

Var. = IndE : Text type = Sp. inf.	−0.742	0.603	−1.231	0.218
Var. = BrE : Text type = Wr. inf.	−0.678	0.676	−1.002	0.316
Var. = IndE : Text type = Wr. inf.	−2.370	0.668	−3.549	< 0.001
Var. = BrE : Text type = Wr. for.	0.610	0.965	0.632	0.527
Var. = IndE : Text type = Wr. for.	−0.777	0.945	−0.822	0.411
Var. = BrE : Coord. = Coord.	1.656	0.603	2.747	< 0.01
Var. = IndE : Coord. = Coord.	1.092	0.572	1.909	0.056
Var. = BrE : Cl. pos. = Initial	2.205	0.534	4.127	< 0.001
Var. = IndE : Cl. pos. = Initial	0.662	0.399	1.658	0.097
Reference values: IndE, Written informal				
Predictor	Estimate	Std. error	Z	p
Intercept	−3.097	0.505	−6.130	< 0.001
Variety = BrE	−0.930	0.727	−1.279	0.201
Variety = SgE	2.856	0.575	4.965	< 0.001
Text type = Spoken informal	0.633	0.472	1.341	0.180
Text type = Spoken formal	0.741	0.508	1.459	0.144
Text type = Written formal	0.135	0.686	0.197	0.844
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.177	0.444	11.670	< 0.001
Clause position = Initial	2.183	0.330	6.611	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239

Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp.inf.	−0.845	0.628	−1.345	0.178
Var. = SgE : Text type = Sp. inf.	−1.628	0.582	−2.796	< 0.01
Var. = BrE : Text type = Sp. for.	−1.692	0.726	−2.330	< 0.05
Var. = SgE : Text type = Sp. for.	−2.370	0.668	−3.550	< 0.001
Var. = BrE : Text type = Wr. for.	−0.305	0.938	−0.325	0.745
Var. = SgE : Text type = Wr. for.	−1.593	0.914	−1.743	0.081
Var. = BrE : Coord. = Coord.	0.565	0.606	0.932	0.351
Var. = SgE : Coord. = Coord.	−1.091	0.579	−1.909	0.056
Var. = BrE : Cl. pos. = Initial	1.544	0.581	2.658	< 0.01
Var. = SgE : Cl. pos. = Initial	−0.661	0.399	−1.658	0.097
Reference value: SgE, Written informal				
Predictor	Estimate	Std. error	Z	p
Intercept	−0.242	0.336	−0.720	0.472
Variety = BrE	−3.786	0.637	−5.939	< 0.001
Variety = IndE	−2.856	0.575	−4.963	< 0.001
Text type = Spoken informal	−0.995	0.352	−2.824	< 0.01
Text type = Spoken formal	−1.629	0.431	−3.778	< 0.001
Text type = Written formal	−1.458	0.629	−2.319	< 0.05
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	4.086	0.417	9.797	< 0.001
Clause position = Initial	1.528	0.237	6.425	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001

Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. inf.	0.784	0.540	1.451	0.147
Var. = IndE : Text type = Sp. inf.	1.628	0.582	2.795	< 0.01
Var. = BrE : Text type = Sp. for.	0.678	0.676	1.002	0.316
Var. = IndE : Text type = Sp. for.	2.370	0.668	3.549	< 0.001
Var. = BrE : Text type = Wr. for.	1.288	0.898	1.434	0.151
Var. = IndE : Text type = Wr. for.	1.593	0.914	1.742	0.081
Var. = BrE : Coord. = Coord.	1.656	0.603	2.747	< 0.01
Var. = IndE : Coord. = Coord.	1.092	0.572	1.909	0.056
Var. = BrE : Cl. pos. = Initial	2.205	0.534	4.127	< 0.001
Var. = IndE : Cl. pos. = Initial	0.662	0.399	1.658	0.097

Table 49. Further ‘Coordination’ contrasts

<i>Reference values: IndE, Coordination</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	2.713	0.527	5.150	< 0.001
Variety = BrE	−1.210	0.718	−1.685	0.092
Variety = SgE	0.136	0.688	0.198	0.843
Text type = Spoken formal	0.108	0.432	0.249	0.803
Text type = Written informal	−0.633	0.472	−1.341	0.180
Text type = Written formal	−0.499	0.668	−0.746	0.456
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05

Coordination = No coordination	−5.177	0.444	−11.671	< 0.001
Clause position = Initial	2.183	0.330	6.611	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. for.	−0.848	0.671	−1.264	0.206
Var. = SgE : Text type = Sp. for.	−0.742	0.603	−1.230	0.219
Var. = BrE : Text type = Wr. inf.	0.845	0.628	1.345	0.179
Var. = SgE : Text type = Wr. inf.	1.628	0.583	2.795	< 0.01
Var. = BrE : Text type = Wr. for.	0.539	0.930	0.580	0.562
Var. = SgE : Text type = Wr. for.	0.035	0.902	0.039	0.969
Var. = BrE : Coord. = No coord.	−0.565	0.606	−0.932	0.351
Var. = SgE : Coord. = No coord.	1.092	0.572	1.909	0.056
Var. = BrE : Cl. pos. = Initial	1.544	0.581	2.658	< 0.01
Var. = SgE : Cl. pos. = Initial	−0.661	0.399	−1.658	0.097
Reference value: SgE, Coordination				
Predictor	Estimate	Std. error	Z	p
Intercept	2.850	0.517	5.509	< 0.001
Variety = BrE	−1.346	0.709	−1.898	0.058
Variety = IndE	−0.136	0.687	−0.198	0.843
Text type = Spoken formal	−0.634	0.421	−1.506	0.132
Text type = Written informal	0.995	0.352	2.825	< 0.01
Text type = Written formal	−0.463	0.625	−0.741	0.459

Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = No coordination	−4.086	0.417	−9.798	< 0.001
Clause position = Initial	1.528	0.237	6.425	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. for.	−0.106	0.665	−0.159	0.874
Var. = IndE : Text type = Sp. for.	0.742	0.603	1.231	0.218
Var. = BrE : Text type = Wr. inf.	−0.784	0.540	−1.451	0.147
Var. = IndE : Text type = Wr. inf.	−1.628	0.582	−2.795	< 0.01
Var. = BrE : Text type = Wr. for.	0.504	0.899	0.560	0.575
Var. = IndE : Text type = Wr. for.	−0.035	0.902	−0.039	0.969
Var. = BrE : Coord. = No coord.	−1.656	0.603	−2.748	< 0.01
Var. = IndE : Coord. = No coord.	−1.092	0.578	−1.909	0.056
Var. = BrE : Cl. pos. = Initial	2.205	0.534	4.127	< 0.001
Var. = IndE : Cl. pos. = Initial	0.662	0.399	1.658	0.097

Table 50. Further ‘Clause position’ contrasts

<i>Reference values: IndE, Initial</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−0.281	0.332	−0.845	0.398

Variety = BrE	−0.231	0.416	−0.554	0.579
Variety = SgE	0.566	0.396	1.430	0.153
Text type = Spoken formal	0.108	0.432	0.249	0.803
Text type = Written informal	−0.633	0.472	−1.341	0.180
Text type = Written formal	−0.499	0.668	−0.746	0.456
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.177	0.444	11.670	< 0.001
Clause position = Non-initial	−2.183	0.330	−6.611	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. for.	−0.848	0.671	−1.264	0.206
Var. = SgE : Text type = Sp. for.	−0.742	0.603	−1.230	0.219
Var. = BrE : Text type = Wr. inf.	0.845	0.628	1.345	0.179
Var. = SgE : Text type = Wr. inf.	1.628	0.583	2.795	< 0.01
Var. = BrE : Text type = Wr. for.	0.539	0.930	0.580	0.562
Var. = SgE : Text type = Wr. for.	0.035	0.902	0.039	0.969
Var. = BrE : Coord. = Coord.	0.565	0.606	0.932	0.351
Var. = SgE : Coord. = Coord.	−1.091	0.579	−1.909	0.056
Var. = BrE : Cl. pos. = Non-ini.	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Non-init.	0.661	0.399	1.658	0.097
Reference value: SgE, Initial				

Predictor	Estimate	Std. error	Z	p
Intercept	0.285	0.291	0.980	0.327
Variety = BrE	−0.797	0.386	−2.064	< 0.05
Variety = IndE	−0.566	0.396	−1.430	0.153
Text type = Spoken formal	−0.634	0.421	−1.506	0.132
Text type = Written informal	0.995	0.352	2.825	< 0.01
Text type = Written formal	−0.463	0.625	−0.741	0.459
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	4.086	0.417	9.797	< 0.001
Clause position = Non-initial	−1.522	0.237	−6.425	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.367	< 0.001
Priming = NP/No priming	0.218	0.164	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = BrE : Text type = Sp. for.	−0.106	0.665	−0.159	0.874
Var. = IndE : Text type = Sp. for.	0.742	0.603	1.231	0.218
Var. = BrE : Text type = Wr. inf.	−0.784	0.540	−1.451	0.147
Var. = IndE : Text type = Wr. inf.	−1.628	0.582	−2.795	< 0.01
Var. = BrE : Text type = Wr. for.	0.504	0.899	0.560	0.575
Var. = IndE : Text type = Wr. for.	−0.035	0.902	−0.039	0.969
Var. = BrE : Coord. = Coord.	1.656	0.603	2.747	< 0.01
Var. = IndE : Coord. = Coord.	1.092	0.572	1.909	0.056
Var. = BrE : Cl. pos. = Non-ini.	−2.205	0.534	−4.127	< 0.001

Var. = IndE : Cl. pos. = Non-ini.	−0.662	0.399	−1.658	0.097

Table 51. Further ‘Accessibility’ contrasts

<i>Reference value: Intermediate</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−5.593	0.687	−8.137	< 0.001
Variety = IndE	1.775	0.644	2.755	< 0.01
Variety = SgE	3.002	0.610	4.920	< 0.001
Text type = Spoken formal	−0.740	0.516	−1.435	0.151
Text type = Written informal	0.211	0.419	0.503	0.615
Text type = Written formal	0.041	0.661	0.062	0.951
Accessibility = High	0.208	0.295	0.704	0.481
Accessibility = Low	−0.933	0.569	−1.639	0.101
Coordination = Coordination	5.742	0.483	11.889	< 0.001
Clause position = Initial	3.727	0.487	7.654	< 0.001
Priming = Omitted pronoun	2.260	0.355	6.368	< 0.001
Priming = NP/No priming	0.218	0.165	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239
Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = IndE : Text type = Sp. for.	0.848	0.671	1.264	0.206
Var. = SgE : Text type = Sp. for.	0.106	0.665	0.159	0.873
Var. = IndE : Text type = Wr. inf.	−0.845	0.628	−1.345	0.179
Var. = SgE : Text type = Wr. Inf.	0.784	0.540	1.451	0.147

Var. = IndE : Text type = Wr. for.	−0.539	0.930	−0.580	0.562
Var. = SgE : Text type = Wr. for.	−0.504	0.900	−0.560	0.576
Var. = IndE : Coord. = Coord.	−0.565	0.606	−0.932	0.351
Var. = SgE : Coord. = Coord.	−1.656	0.603	−2.747	< 0.01
Var. = IndE : Cl. pos. = Initial	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Initial	−2.205	0.534	−4.127	< 0.001

Table 52. Further ‘Priming’ contrasts

<i>Reference value: NP/No priming</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−5.375	0.681	−7.884	< 0.001
Variety = IndE	1.775	0.644	2.755	< 0.01
Variety = SgE	3.002	0.610	4.920	< 0.001
Text type = Spoken formal	−0.740	0.516	−1.435	0.151
Text type = Written informal	0.211	0.419	0.503	0.615
Text type = Written formal	0.041	0.661	0.062	0.951
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.742	0.483	11.889	< 0.001
Clause position = Initial	3.727	0.487	7.654	< 0.001
Priming = Omitted pronoun	2.042	0.357	5.714	< 0.001
Priming = Overt pronoun	−0.218	0.165	−1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Modal aux.	−0.403	0.343	−1.176	0.239

Verb class = Non-modal aux.	−1.146	0.304	−3.771	< 0.001
Var. = IndE : Text type = Sp. for.	0.848	0.671	1.264	0.206
Var. = SgE : Text type = Sp. for.	0.106	0.665	0.159	0.873
Var. = IndE : Text type = Wr. inf.	−0.845	0.628	−1.345	0.179
Var. = SgE : Text type = Wr. Inf.	0.784	0.540	1.451	0.147
Var. = IndE : Text type = Wr. for.	−0.539	0.930	−0.580	0.562
Var. = SgE : Text type = Wr. for.	−0.504	0.900	−0.560	0.576
Var. = IndE : Coord. = Coord.	−0.565	0.606	−0.932	0.351
Var. = SgE : Coord. = Coord.	−1.656	0.603	−2.747	< 0.01
Var. = IndE : Cl. pos. = Initial	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Initial	−2.205	0.534	−4.127	< 0.001

Table 53. Further ‘Verb class’ contrasts

<i>Reference value: Non-modal auxiliary</i>				
Predictor	Estimate	Std. error	Z	p
Intercept	−5.385	0.620	−8.686	< 0.001
Variety = IndE	1.775	0.644	2.755	< 0.01
Variety = SgE	3.002	0.610	4.920	< 0.001
Text type = Spoken formal	−0.740	0.516	−1.435	0.151
Text type = Written informal	0.211	0.419	0.503	0.615
Text type = Written formal	0.041	0.661	0.062	0.951
Accessibility = Intermediate	−0.208	0.295	−0.704	0.481
Accessibility = Low	−1.141	0.529	−2.158	< 0.05
Coordination = Coordination	5.742	0.483	11.889	< 0.001
Clause position = Initial	3.727	0.487	7.654	< 0.001

Priming = Omitted pronoun	2.260	0.355	6.368	< 0.001
Priming = NP/No priming	0.218	0.165	1.324	0.185
Clause type = Subordinate	−1.222	0.201	−6.076	< 0.001
Pronoun = Other	0.527	0.207	2.542	< 0.05
Verb class = Lexical	1.146	0.304	3.771	< 0.001
Verb class = Modal aux.	0.743	0.413	1.799	0.072
Var. = IndE : Text type = Sp. for.	0.848	0.671	1.264	0.206
Var. = SgE : Text type = Sp. for.	0.106	0.665	0.159	0.873
Var. = IndE : Text type = Wr. inf.	−0.845	0.628	−1.345	0.179
Var. = SgE : Text type = Wr. Inf.	0.784	0.540	1.451	0.147
Var. = IndE : Text type = Wr. for.	−0.539	0.930	−0.580	0.562
Var. = SgE : Text type = Wr. for.	−0.504	0.900	−0.560	0.576
Var. = IndE : Coord. = Coord.	−0.565	0.606	−0.932	0.351
Var. = SgE : Coord. = Coord.	−1.656	0.603	−2.747	< 0.01
Var. = IndE : Cl. pos. = Initial	−1.544	0.581	−2.659	< 0.01
Var. = SgE : Cl. pos. = Initial	−2.205	0.534	−4.127	< 0.001

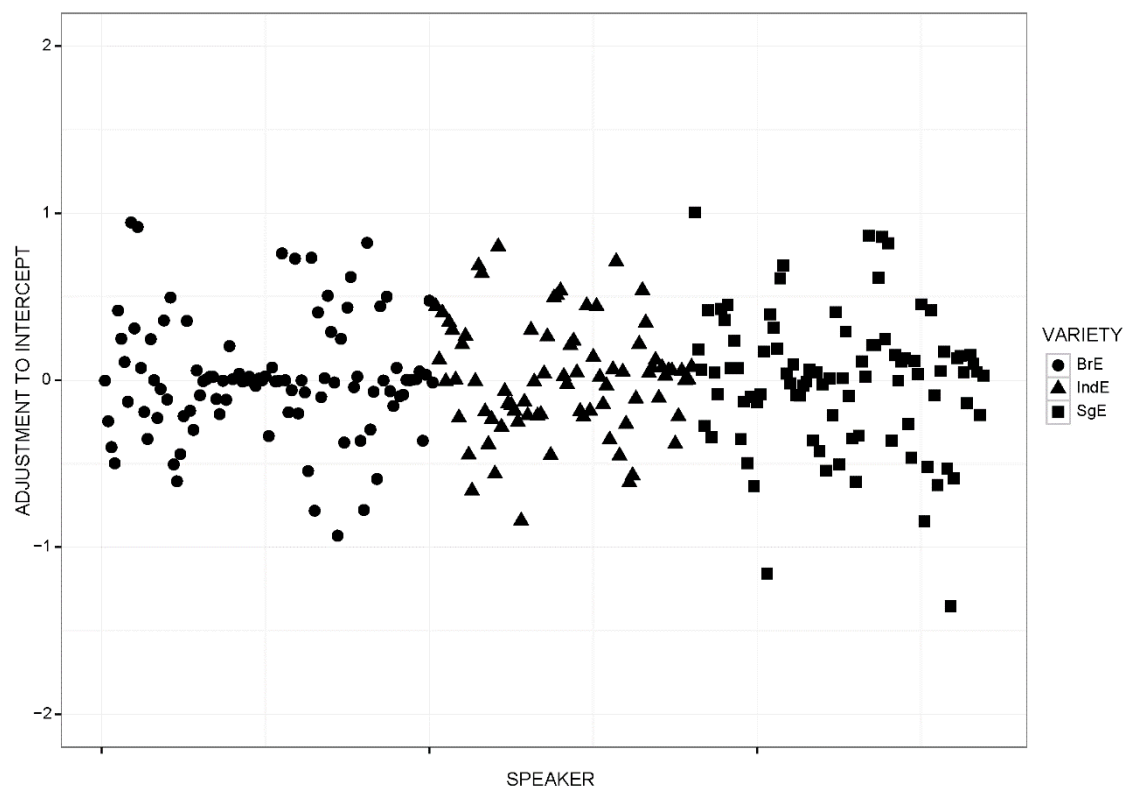


Figure 37. Speaker preferences across varieties